Universitat Autònoma de Barcelona

Dipòsit digital de documents de la UAB

This is the **published version** of the bachelor thesis:

Morató Catafal, Alba; Doval Diéguez, Eduardo, dir. Analysis of aberrant response patterns in educational multiple choice tests. 2018. 51 pag. (954 Grau en Psicologia)

This version is available at https://ddd.uab.cat/record/200028

under the terms of the BY-NC-ND license

Analysis of aberrant response patterns in educational multiple choice tests

Author: Alba Morató Catafal
Supervisor: Eduardo Doval Diéguez

September 25, 2018

Abstract

The objective of this project has consisted on providing empirical evidence of the validity of individual scores in an evaluation test by performing an aberrant pattern analysis of the results of the PIR exam of 2005. The tool used has been person-fit analysis, which allows to detect anomal patterns in the individual responses to the test that might imply a bias in the score of a test to infer a trait level. The person-fit indices computed have been H_t, $C*$ and lz (with their respective cutoff values). These indices have been computed for nine sub-datasets into which the data has been splitted, in order to accomplish one of the assumptions of the models. For the same reason, some of its items have had to be erased and an IRT model (1, 2 or 3 parameters) has had to be chosen for each of the sub-datasets. After that, the individuals who had obtained the best marks in the test have been analysed. For that, their responses to certain of the sub-datasets that had been flagged as aberrant have been compared to simulated data in order to visually identify the type of aberrant patern commited by them. It has been found that the majority of them could actually be classified as normal, some may have been lucky guessers and a few present cheating patterns. However, this information is not a firm conclusion but consists on indicators that should be complemented by further information such as interviews.

Resumen

El objetivo de este proyecto ha sido proporcionar evidencia empírica de la validez de las puntuaciones individuales del examen PIR de 2005 mediante un análisis de patrones atípicos de respuesta de sus resultados. La herramienta usada ha sido el análisis *person-fit*, que permite detectar patrones atípicos en las respuestas de los individuos al test, lo cuál puede implicar un sesgo en la puntuación obtenida en esta, de la cual se infiera un nivel de rasgo, en el caso del PIR conocimientos en psicología. Los índices calculados han sido el H_t, el $C*$ y el lz (con sus respectivos puntos de corte). Estos índices han sido calculados para nueve sub bases de datos en las que se ha dividido la base de datos inicial, para cumplir con una de las asunciones del modelo. Por la misma razón, algunos de los items han tenido que ser eliminados y un modelo de IRT (1, 2 o 3 parámetros) ha sido elegido para cada una de las sub bases de datos. Después de esto, los individuos que habían obtenido las mejores notas en el examen han sido analizados. Para eso, sus respuestas a algunas de las areas del test que habían estado marcado como atípicas han sido comparadas con datos simulados para así idenficar de manera visual el tipo de patrón atípico de respuesta cometido en cada caso. Así, se encontró que la mayoría de estos podía ser en realidad clasificado como normal, algunos parecían haber tratado de adivinar las respuestas y finalmente, unos pocos parecían mostrar indicios de haber copiado. De todas formas, esta información no constituye una conclusión firme, sin'o que se trata de indicadores que deberían ser contrastados y complementados con más información, como entrevistas.

Resum

L'objectiu d'aquest projecte ha estat proporcionar evidència empírica de la validesa de les puntuacions individuals de l'examen PIR de 2005 mitjançant un anàlisi de

patrons atípics de resposta dels seus resultats. L'eina emprada ha estat l'anàlisi *person-fit*, que permet detectar patrons atípics de resposta dels individus al test, el qual pot implicar un biaix en la puntuació obtinguda en aquest, de la qual s'infereix un nivell de tret, en el cas del PIR de coneixements en psicologia. Els índexs calculats han estat el H_t, el $C*$ i el lz (amb els seus respectius punts de tall). Aquests índexs han estat calculats per a nou sub bases de dades en les que s'ha dividit la base de dades inicial, per tal de complir amb les assumpcions del model. Per la mateixa raó alguns dels ítems han hagut de ser eliminats i un model d'IRT (1, 2 o 3 paràmetres) ha estat ajustat per a cadascuna de les sub bases de dades. Després d'això, els individus que havien obtingut les millors notes a l'examen han estat analitzats. Per això les seves respostes en algunes de les àrees del test que havíen estat marcades com a atípiques han estat comparades amb dades simulades per a així identificar de manera visual el tipus de patró atípic de resposta comés en cada cas. Així, s'ha trobat que la majoría d'aquests podía ser en realitat classificat com a normal, alguns semblava haver tractar d'endevinar les respostes i finalment, molt pocs, semblaven presentar indicis d'haver copiat. De totes maneres cal tenir en compte que la informació obtinguda no constitueix una conclusió ferma sinó que es tracta d'indicadors, que hauríen de ser contrastats i complementats amb més informació com entrevistes amb els estudiants.

Contents

1 Introduction

In education, multiple choice tests are often used in order to evaluate the level of knowledge of students. However, it is known that total scores do not always reflect the proficiency level tests intend to measure. The answers provided by test takers may be biased due to factors unrelated to the trait of interest. For example, cheating or item-preknowledge may inflate exam scores whereas inattention or guessing may deflate them. It is important to be able to detect whether inferences from scores are invalid, that is, whether test scores are biased and therefore not indicative of the true latent trait being measured (Tendeiro, Meijer & Niessen, 2016).

In order to detect anomalies in individual responses to a test, one can perform an analysis of aberrant patterns of the answers of the test takers. This kind of analysis is also called person-fit analysis and it consists on detecting when a test performance of an individual deviates from a behaviour that is exhibited by most examinees and that differs from normal or common item score patterns (Meijer & Sijtsma, 1995b).

As an illustration, let us consider an examinee that takes an exam of, say, 100 items. Of these 100 items, 35 are defined as being easy, 35 of medium difficulty and 30 difficult. If the examinee gets a 70/100 one would expect he or she answered right the majority of the easiest and medium difficulty items and that the ones he or she failed were mainly among the most difficult ones. This would be a normal response pattern. But let us suppose that in this exam the easiest questions are related to a specific topic and that our examinee has not studied it, he/she will then have answered right some of the most difficult questions without having answered right the easiest ones. Assuming that this strategy is unusual among the other examinees of similar ability and given a test model that assumes that the probability of obtaining the correct answer on any item increases with the ability, his pattern would be classified as atypical or aberrant.

1.1 Objective

Person-fit analysis allows to detect anomal patterns that might imply a bias in the score of a test to infer a trait level. That is why the objective of this study will be providing empirical evidence on the validity of individual scores in an evaluation test by performing an aberrant pattern analysis to real data. The test analysed has been the PIR exam of 2005. PIR is a Spanish exam that allows the ones that obtain the best marks to become Resident Internal Psychologists and become specialists in clinical psychology.

As some test takers will appear as aberrant respondents, it would be interesting to inspect the relationship between being flagged as aberrant respondent and obtaining the PIR internship. This could be checked by maintaining the anonymity, just by knowing how many interships where granted. Another interesting point would be exploring the different patterns between the aberrant respondents in order to check if any of them seem to follow a visible strategy and if it could be improved.

But first of all it is indispensable to understand the different models used in person-fit analysis and learn how to perform them, which will be treated in the following section.

2 Models

In the field of person-fit analysis there are several statistics that allow to detect aberrant response patterns. The majority of these statistics are either based on the Item-Response Theory or are group-based indices (Núñez & López, 2006).

Let us briefly introduce the concepts of the Item-Response Theory (IRT). The IRT was born as a response to the Classical Test Theory in psychology, trying to overcome its problems and provide a different approximation to the description of tests characteristics and its psychometrics by not looking only at the total score, but also at the items individually. The Classical Test Theory and the improvements made by IRT will not be treated here but further information can be found elsewhere (Abad, Olea, Ponsoda & García, 2014).

IRT is based on the item characteristic curve (ICC), which is a curve resulting of plotting the probability that an examinee with a certain ability or level of a trait will give a correct answer to the item as a function of his ability and the psychometric characteristics of the item. This function can be modeled (parametrically or non-parametrically) and describes how the probability of answering the item correctly changes through the different values of the measured trait.

In the parametric case, it is needed to estimate the level of the individual asociated with the measured variable and also up to three psychometric characteristics of each item (difficulty, discriminability and guessing). This information is then translated into the model, frequently a logistic one, that depends on these parameters and models the probability of answering right the item. In the non-parametric case, the difficulty parameter can be estimated, for example, by the proportion of correct answers.

Finally, there are the group-based statistics. These are not based on the IRT or other models but on group characteristics (Meijer & Sijtsma, 1995). In general, these indices tend to classify item score patterns with many correct answers on items that most people of the group have answered wrongly as aberrant.

2.1 Assumptions

In order to compute the person-fit indices some assumptions are made, depending on how the index is constructed and need to be checked before of computing them. Group-based statistics only assume unidimensionality. The IRT parametric indices assume unidimensionality and also local independency. Finally, for the non-parametric IRT models, a third assumption has to be added: latent monotonicity of the item characteristic curves. For

the IRT parametric models, it is also important to check the fit of the model chosen.

Unidimensionality means that the test measures just one dimension, so the probability of answering correctly an item depends on its parameters and the level of the variable being measured on the person, but not on other variables such as intelligence or vocabulary of the person if these are not the aimed traits. This assumption can be checked through multiple tools.

Local independency between the items of a test implies that the answer from a person to one of them, conditioned to his/her latent trait, does not depend on its answer to the other ones. This definition is a way of stating that the latent variable explains why the observed items are related to one another.

Finally, in non-parametric IRT based models, the ICCs are assumed to be monotone nondecreasing in the latent trait (θ), what means that the item step response functions are nondecreasing functions of θ. That is, for two arbitrary fixed values of this trait θ_a and θ_b, and a test with J items:

$$P_j(\theta_a) \leq P_j(\theta_b), \text{ whenever } \theta_a < \theta_b; \quad j = 1, ..., J. \tag{1}$$

2.2 Person-fit indices

Many person-fit statistics have been proposed in the literature. Although the ones being used in this study will be described below, first it will be explained why these and not any others have been chosen. Karabatsos (2003) discusses and compares the most relevant ones and concludes that of the 36 person-fit statistics examined, H_T is the best in identifying aberrant-responding examinees, and C, MCI, and U_3 are some of the second best. Of these mentioned indices, C, MCI and H_T are Group-Based and U_3 is related to IRT non-parametric models. Moreover, although not being included in the study as one of the most remarkable, it has also been included the IRT parametric index l_z, as a revision of the literature has shown to be widely used.

Despite, as it has just been said, they are already implemented in the tool that will be used, let us see how they are obtained and, most importantly, how to interpret these indices:

The Caution index (C) was proposed by Sato (1975) and is a covariance ratio measuring the extent to which an item score pattern deviates from the perfect pattern (Guttman pattern) that would consist on succeeding on all the items up to a certain difficulty, and then failing on all the items above that difficulty. It is:

$$C = 1 - \frac{\text{Cov}(x_n, p)}{\text{Cov}(x_n^*, p)}, \tag{2}$$

where x_n is the 0/1 response vector of individual n, x_n^* is the correspondent Guttman vector, which is obtained by ordering the 0/1 response vector of the individual by placing

the ones in the first positions (after the items being ordered by decreasing proportion-correct score) and the zeros in the lasts and p is the vector of item proportions-correct. Higher values of this index indicate aberrant patterns but it has no upper bound. The lower bound is 0.

Harnisch and Linn (1981) further adapted C by limiting it to a range [0, 1]. This statistic is called C^* or Modified Caution Index (MCI) and is a ratio between two covariances: the covariance of X_n with the perfect pattern and the covariance of X_n with the perfectly-inconsistent pattern. Among J test items, the perfectly inconsistent pattern contains correct responses for only the most difficult items. It can be obtained by:

$$\text{MCI} = \frac{\text{Cov}(x_n^*, p) - \text{Cov}(x_n, p)}{\text{Cov}(x_n^*, p) - \text{Cov}(x_n', p)}, \tag{3}$$

where x_n' is the reversed Guttman vector (errors in the easiest items) containing correct answers for the $J-S$ hardest items, with the smaller proportion-correct values only (being J the total number of items and S the total score obtained $S = \sum_{j=1}^{J} X_j$). MCI ranges between 0 (perfect Guttman vector) and 1 (reversed Guttman vector).

The U_3 statistic (Van der Flier, 1982) is similar to C^*. Suppose that the items are ordered in decreasing proportion-correct score, $p_1 > p_2 > p_3 > ... > p_J$, where J is the number of items. Given a response vector $(X_1, X_2, ..., X_J)$ with total score $S = \sum_{j=1}^{J} X_j$, the U_3 statistic is defined as:

$$U_3 = \frac{\sum_{j=1}^{S} p_j - \sum_{j=1}^{J} p_j}{\sum_{j=1}^{S} p_j - \sum_{j=J-S+1}^{J} p_j}. \tag{4}$$

U_3 varies from 0, for perfect Guttman response vectors, response patterns where the S first items (after being ordered by decreasing proportion-correct score) are ones and the rest are zeros, to 1 for reversed Guttman response vectors.

The H_T statistic was adapted by Sijtsma (1986) from a former statistic introduced by Mokken (1971). It measures the similarity between the n^{th} examinee's response vector X_n with the response vectors of the remaining $N-1$ examinees. It is given by:

$$H_T = \frac{\text{Cov}(X_n, r_{(n)})}{\text{Cov}_{max}(X_n, r_{(n)})}, \tag{5}$$

assuming that the rows of the data matrix are ordered by increasing order of total score S_n ($n = 1, ... , N$) and $r_{(n)}$ is the vector of total item scores computed excluding individual n and the denominator is the maximum covariance given the marginal. H_T takes its maximum value 1 when no respondent with a total score smaller/larger than S_n can answer an item correctly/incorrectly that respondent n has answered incorrectly/correctly, respectively. The possible range of the statistic is [-1, 1]. H_T will be positive when the responses by a person are positively correlated with all the other persons and will be negative when a person is negatively correlated with all the other persons. When person's responses are

random, H_T will be close to zero. Hence, aberrant response behavior is indicated by small values of H_T.

Finally, Drasgow et al. (1985) introduced the l_z index, one of the most used person-fit statistics. This statistic is a unit normal transformation of the index l, which measures the log-likehood fit of an individual's responses X_n with the predictions of an IRT model. The computation of l_z requires that both item and ability parameters are available. The ability parameter estimates are obtained by using the maximum likelihood method. Also, although there can be used up to three parameters for the items, only one will be used here, which will be estimated by maximum likelihood too. Therefore, l_z can be obtained as:

$$l_z = \frac{l - E(l)}{\sqrt{v(l)}}, \tag{6}$$

where

$$l = \sum_{j=1}^{J} [X_{nj}(\log P_{nj1}) + (1 - X_{nj})(\log P_{nj0})] \tag{7}$$

and

$$E(l) = \sum_{j=1}^{J} [P_{nj1}(\log P_{nj1}) + (P_{nj0}(\log P_{nj0})], \tag{8}$$

where X_{nj} is the examinee n's scored response to test item j, P_{nj1} is the probability of a correct ($X_{nj} = 1$) response and therefore P_{nj0} is the probability of an incorrect ($X_{nj} = 0$) response, both predicted by an IRT model, with $P_{nj0} = (1 - P_{nj1})$. Aberrant response behavior is indicated by small values of l_z, that indicate low likelihood.

As the first three described indices, C, MCI and $U3$, have the same basic form, and in order to simplify the study, only one of them will be used. U3 will be used, as its a non-parametric IRT based index and therefore, three different kinds of indexs will be used through the study.

3 Methods

The database that will be analysed is from the PIR exam of 2005. Data was provided by the Spanish Ministerio de Sanidad y Consumo. As it has already been introduced, PIR is a multiple choice exam done in Spain that allows the ones that obtain the best marks to become Resident Internal Psychologists and become specialists in clinical psychology.

The data being used in this study includes 2057 rows, representing the individuals that have taken the test and 260 columns, one for every item of the test. After applying the answer key to the raw results of the test takers, the data only has 0s, for wrong answers, 1s, for correct answers and NAs for not responded items. For the purpose of the study NAs have been coded as wrong answers, as one can supose that someone who has left a question blank is because he/she did not know its answer. Regarding ethical

concerns, individuals will be associated to an identifier code not directly relating them to the persons to whom the data belongs.

About the content of the data, the PIR exam includes questions from different fields on psychology. In (Moreno, Martínez. & Muñiz, 2011) these items were classified in the following 9 areas of psychology:

1. Psychopathology
2. Therapies and treatments
3. Psychodiagnostic and conductual evaluation
4. Personality and diferential psychology
5. Basic processes and history
6. Psychometrics, statistics and methods
7. Social and organizational psychology
8. Evolutive and educational psychology
9. Psychobiology and psychophysiology

Knowing this classification of the items it can be pressuposed that taking all of the items as a block will lead into problems of unidimensionality, as more than one dimension is evaluated in the test so, although this will be checked, the test will be analysed by subtests based on these areas and in the end, the results will be put together.

In order to work with the data the R software has been used.

To start with, a brief descriptive analysis of the data will be performed, including the proportions for all the possible response categories for each item and the frequencies of all possible total scores of the test.
Then, the three mentioned assumptions will be checked.

To study the unidimensionality the tool used will be the one proposed by Drasgow and Lissak (1983), which is implemented in the `unidimTest` function of the R `ltm` package. Its objective is examining the latent dimensionality of dichotomously scored item responses. The statistic used for testing unidimensionality is the second eigenvalue of the tetrachoric correlations matrix of the dichotomous items. A Monte Carlo procedure is used to approximate the distribution of this statistic under the null hypothesis of unidimensionality. The value of the statistic (i.e., the second eigenvalue) for the original data-set is denoted T_{obs}. Then the p-value is approximated according to the formula:

$$p - val = 1 + \frac{\sum_{b=1}^{B} I(T_b \leq T_{obs})}{(1 + B)} \tag{9}$$

Where $I(.)$ denotes the indicator function, and T_b denotes the value of the statistic in the bth data-set of the Monte Carlo procedure (Rizopoulos, 2017). This p-value would allow us to contrast the null hypothesis of unidimensionality.

Then local independence should be checked. To do so it has been adapted to R the MODFIT tool (Stark, 2007), originally in Excel[1]. The tool consists on examining the model fit of individual items, item pairs and item triples by performing χ^2 tests. χ^2/df values larger than 3 in the item pairs and triples χ^2/df indicate a violation of local independence. The tool computes de adjusted χ^2 degrees of freedon ratios (χ^2/df) introduced by Drasgow et al. (1955). These χ^2 statistics are based on expected frequencies that depend on the estimated item parameters and the distribution of θ. The unadjusted statistic for item j is given by:

$$\chi^2_j = \sum_{z=0}^{C} \frac{(O_{jz} - E_{jz})^2}{E_{jz}}, \tag{10}$$

with

$$E_{jz} = N \int P_{jz}(\theta)\phi(\theta)d\theta$$

Where O_{jz} is the observed frequency the answer z (0 or 1) for the item j and $\phi(\theta)$ is the standard normal density. The equation above applies to single items. The formula is easily extendible to pairs and triplets of items, computing in these cases P_{jz} under the assumption of independence. As an heuristic, values of χ^2/df larger than 3 are indicative of model misfit or in this case, absence of local independence.

For the assumption of non decreasing monotonicity it will be used the `check.monotonicity` function from the R `mokken` package. Junker and Sijtsma (2000) showed that for dichotomous items latent monotonicity implies manifest monotonicity. Manifest monotonicity is an observable property of the test data, and is defined as:

$$P(X_j \geq x \mid R_{-j} = s) \geq P(X_j \geq x \mid R_{-j} = r) \text{ for all } j, x, s > r, \tag{11}$$

being X_j the score on item j (0 or 1), R_{-j} the rest score, defined as $R_{-j} = (\sum_{j=1}^{J} X_j) - X_j$. A practical issue to take into account is that some violations of manifest monotonicity may be too small to be relevant. Therefore, only violations greater than a minimum value (usually set as 0.03) are reported and for each reported violation a significance test at level $\alpha = 0.05$ (without Bonferroni correction) is computed (Molenaar and Sijtsma, 2000), so only significant violations will be taken into account.

Finally, the $U3$, H_T and lz indices will be computed for each individual, with the R `PerFit` package, in order to identify aberrant patterns within the test takers. From these indices a cutoff is needed in order to decide which individuals are flagged as aberrant respondants. To do so 1000 repetitions of the model-fitting item response vectors are

[1]Special thanks to J. Tendeiro, who has adapted the code by personal requirement.

generated based on the proportion of respondents per answer category. This allows computing a sample of 1000 values of the person fit statistic corresponding to model-fitting item response patterns. A bootstrap procedure is then used to approximate the sampling distribution of the quantile of level 0.05 for the most extreme types of person fit statistics (depending on the statistic this could be the upper or lower tails), based on 1000 resamples. The cutoff (and its standard error) is given by the median (standard deviation) of this bootstrap distribution. This procedure needs to be applied for every index in every dataset of every subtest.

An individual will be categorized as aberrant respondent if he/she has been marked as aberrant by any of the indices used. Afterwards, the marks obtained by all the individuals will be computed by following the official criterion. Knowing that that year there were 89 positions available, we can infer the 89 best marks as the ones that obtained the internship and analyse which of those individuals' responses have been marked as aberrant. From those, it would be interesting to see which kind of aberrant pattern they have commited in order to see if any of them could actually not have the knowdledge infered by the mark (e.g. cheaters or lucky guessers). To do so, a good tool would be interviewing each of them, but as in this case it is not possible, some alternative tools are introduced here in order to try to understand these response patterns. A possibility being explored in this project is, regarding an aberrant response pattern, simulating different possible patterns taking into account the person's estimated ability and the difficulty of the items of the test. The patterns simulated would be: normal respondents, cheaters, random, careless and creative respondents, followin the method in Karabatsos (2003). Then, these patterns could be plotted and compared to the original pattern of the person being analysed in order to see if it looks like any of these and therefore, assume this person can be classified as that type of respondent.

4 Results

To start with, a brief descriptive analysis of the data has been done. Some plots can be found in appendix A. Starting with the missing data, it can be seen in Figure 6 that there are some individuals with 250 items not responded (out of 250). These individuals that have not answered any of the items will be removed from the data as they not provide any information. Also, from now on, missing data will be considered as not known items, being this missing responses categorized as wrong answers (0s). Also, the division made by the content of the items (by areas of study of psychology) will be used to inspect and analyse the data. We can start forming a first idea about the difficulty of the items by the proportion of correct answers on each of them (proportion of 1s), that can be seen in Figures 8 and 9: there seem to be items with a wide range of difficulties and with variable difficulties in all the areas. Also from this plot, it seems that the areas where most of the items come from are Psychopathology (54 items) and Therapies and treatments (41), while Personality and diferential psychology is the one with least items (13), with the other areas having around 20 items each. About the punctuations obtained, from Figure 10 it seems that in some areas the tendence is to get higher scores, like in Psychopathology

and psychodiagnostic; whilst in others as Evolutive and educational psychology and Psychobiology and psychophysiology the mean seems to be lower; in psychometrics, statistics and methods it seems that aproximately all intervals of scores occur with the same frequency and finally the rest seem to present an approximately normal distribution.

After this descriptive analysis of the data the assumptions of unidimensionality, local independency and non-decreasing monotonicity of the ICC should be checked among the data. The first assumption that will be checked will be the non-decreasing monotonicity as it affects the items and could imply the supression of some of them. From all of the 250 items of the test, there are 18 items do not discriminate well. These items come from all the different subsets of the test except the psychometrics, statistics and methods one. To see the item step response function of these items and to which subtest do each of them pertain check appendix B. These items will be removed from the analysis, as deleting them will not cause any problem regarding the size of the sample, and as said in Sijtsma and Molenaar (2002) they might disturb the ordering of the respondents by means of the number of correct answers. If a pilot study had been done for the test they could had been advised to be removed from its final version.

After, local independence should be checked. For that, the MODFIT procedure has been used, and for each subset, the model with lower mean values of the statistics has been the one used to fit the data. While chosing the model the values for single items have also been taken into account, in order to choose a model with a good fit. To illustrate it, let us see the results of the MODFIT procedure for the items regarding Psychopathology in Tables 1, 2, 3. The three tables represent the fit for the IRT models of 1, 2 or 3 parameters. As it can be seen, for the 1PL model all of the singlets adjusted χ^2 values are below 3, which indicate a good fit of the model, but the same cannot be said for the values of doublets and triplets, therefore let us check the 2 parameter model. This model present better results for doublets and triplets, at it indicates the mean for this values, that is lower than for the 1PL model, nevertheless there is one value for the singlets that is higher, although anyways it keeps being lower than 3. Finally, by checking the 3 parameter model it can be seen that this would be the best model: the mean for the singlets adjusted χ^2 is of 0, which indicates a good fit and also, the adjusted chi^2 values for doublets and triplets are the lower among the three models, as it is reflected in the means of these values (0.35 and 0.57), which indicated that by using this model, local independence assumption is controlled. For the other areas the same procedure has been followed; its MODFIT procedure results can be checked in appendix C (Note: In some cases the 3PL model has not converged, in this case the results of this model have not been provided nor considered) and the chosen models can be checked in Table 4.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	50	0	0	0	0	0	0	0.00	0.00
Doublets	461	106	85	85	47	94	347	6.25	10.25
Triplets	1768	1764	1755	1832	1690	2950	7841	8.04	8.11

Table 1: Results from the analysis of the local independence and the fit of the 1PL IRT model for the Psychopathology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	49	1	0	0	0	0	0	0.06	0.20
Doublets	1027	102	48	25	10	10	3	0.49	1.10
Triplets	13478	3740	1455	551	212	108	56	0.84	1.08

Table 2: Results from the analysis of the local independence and the fit of the 2PL IRT model for the Psychopathology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	50	0	0	0	0	0	0	0.00	0.01
Doublets	1084	84	31	12	7	3	4	0.35	0.94
Triplets	15552	2871	805	233	72	22	45	0.57	0.85

Table 3: Results from the analysis of the local independence and the fit of the 3PL IRT model for the Psychopathology data.

Next, unidimensionality is checked (figures in appendix D). To do so, the best IRT model chosen (in the previous step) for each dataset is used. As it could be suspected, it seems the unidimensionality assumption is not accomplished for all the data as a unique test, as it is actually composed of different dimensions. By dividing it by the 9 areas of the exam the unidimensionality test keeps pointing a rejection of the hypothesis except on the case of the Personality and diferential psychology subtest. Nevertheless, it can be seen that the by dividing in subtests the data its structure seems to be more unidimensional than all together data, and after analysing it one by one it can be said, following the criterion on Drasgow and Lissak (1983), that these datasets present basic unidimensionality, which will be accepted to perform the analysis.

Finally, a brief summary of all the assumption checking can be seen in Table 4.

After all the checkings and changes in the data, the chosen indices ($C*$, Ht and lz) can be computed. Let us see how do the indices distribute among the data: As it can be seen in apendix E, in aproximately the 90% of the cases the respondants' answers are not flagged as aberrant. When they are, there does not seem to exist a clear pattern: there is a nearly equal percentage of cases being flagged by one, two or the three indices computed (around the 3%). $C*$ and H_t seem to correlate, as they often classify the same patterns as aberrant. Also, lz is the index that seems to classify more cases as aberrant when the others do not. Hence, it seem that the three indeces complement each other, as they are computed in different ways and they do not flag the same cases.

	MND items	Basic unidim	Model
Psychopathology	214, 215, 234, 254	✓	3PL
Therapies and treatments	149, 188	✓	2PL
Psychodiagnostic	197	✓	3PL
Personality and diferential psychology	191	✓	3PL
Basic processes and history	4, 28	✓	3PL
Psychometrics, statistics and methods		✓	3PL
Social and organizational psychology	128, 133	✓	2PL
Evolutive and educational psychology	75, 176, 184	✓	3PL
Psychobiology and psychophysiology	19, 47, 57	✓	3PL

Table 4: Results from the assumption checking in each of the sub-datasets: first column show the items that do not accomplish the non-decreasing monotonicity assumptions and therefore have been removed; second column shows the check of basic unidimensionality and third one indicates the model chosen among 1, 2 or 3 parameter logistic model (1PL, 2PL, 3PL) after checking the goodness of fit and the indicators of local independence.

The marks obtained by all the individuals have been computed by following the official criterion: wrong responses count as -3, right ones as 1, and not answered ones as 0, and then the sum of these has to be multiplied by a correction factor, obtained by dividing 90 by the mean of the 10 best marks. These marks have been plotted for all the individuals, seeing now the distribution of the results obtained by the respondents. In the plot it has been superposed the proportion of flagged responses as aberrant for each of the areas of the test (see Figure 1) it can be seen that there are a really high proportions of individuals presenting aberrant patterns in the left tail, which could be explained because in some of these groups there is only one individual, so the proportions can only get the values 0 or 1. In the right tail there is also a high proportion individuals flagged as aberrants. Knowing that that year there were 89 positions available, we can infer the 89 best marks as the ones that obtained them (although actually the mark on the exams ponderates 0.9 for the final punctuation, as it also is needed to take into account the academic record). Of these 89 cases, nearly the 50% did not present any aberrant response pattern in any of the parts of the test. About the other 50% of these respondants, the 40% presented an aberrant response pattern in one of the areas of the test, the 5% presented two aberrant response patterns and also the 5% presented three.

Now, the important part would be analysing one by one the flagged cases, in this case it has been given more importance to the ones that have obtained the position, in order to see if any of the ones that obtained it could actually not have the knowdledge infered by the mark (e.g. cheaters or lucky guessers). Also, it would be interesting analysing the people who nearly obtained the position, but did not because they made an aberrant pattern (e.g. creative or careless respondents), as some advice could be provided to them in order to improve for the next time. To do so, a good tool would be interviewing each of them, but as in this case it is not possible, some alternative tools are introduced here in order to try to understand these response patterns. A possibility being explored in this project is, regarding an aberrant response pattern, simulating different possible patterns taking into account the person's estimated ability and the difficulty of the items of the

test. Then, for example, patterns for cheaters, random respondents, etc. could be plot and compared to this person's pattern in order to see if it looks like any of these and therefore, assume this person can be classified as that type of responant.

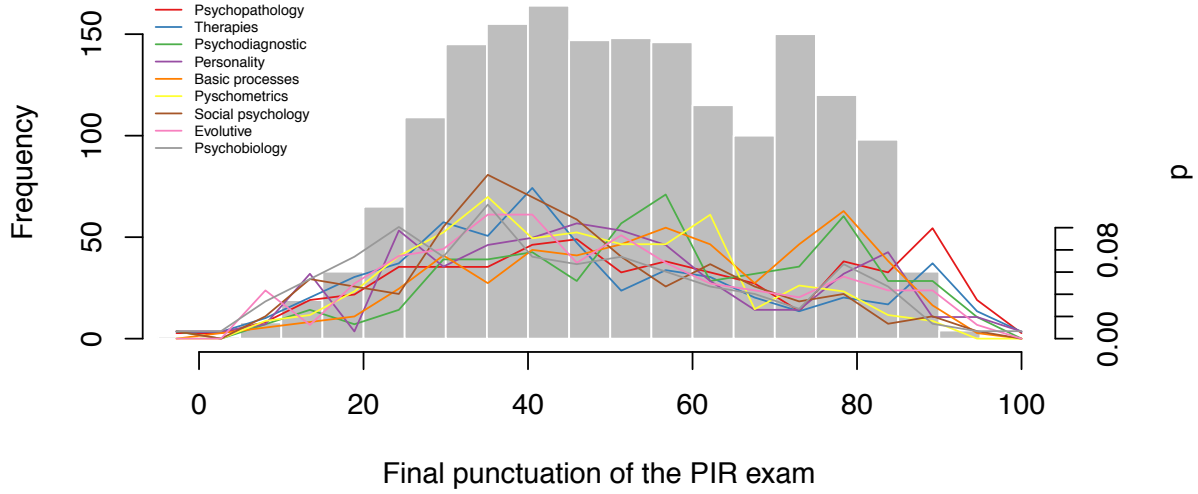


Figure 1: Histogram of the marks (up to 100) obtained by the respondents of the test. The axis of the right shows the proportion of flagged respondants as aberrant and corresponds to the lines over the plot. There is one line for each of the areas in which the test has been divided.

For example, the 380th case of our database has been classified as aberrant respondant, for the Psychopathology subset of items. This person's responses are shown in Figure 2. As it is explained in the caption, in the x axis the items are shown, ordered by its estimated difficulty, and this difficulty is reported in the y axis. The vertical line shows the number of correct responses given by the person. Each rectangle of the background is a part of the test, having been this divided into three tertiles of same number of items aproximately: the first area (green) would correspond to the easiest items; the last one (red) would correspond to the most difficult ones and the one in the middle (orange) would correspond to intermediate difficulty items. Green points over the line show an item has been answered correctly and red ones are wrong answered items. It would be expected, for normal respondants, to have correct answers (green points) on the left side of the vertical line, while in the right side of the line it would be expected to have more red points (wrong answers). This information is summarized in Figure 3, where, as it is also explained, each rectangle is a part of the test. The first and last ones correspond to the first and last tertiles while the central ones correspond to the second tertile, divided by the mark obtained by the person (the vertical line). In each of the rectangles the green coloured circles indicate the proportion of correct answers obtained, while the red ones indicate the proportion of wrong answers for that part of the test. The outter circles are

an indicator of the total, but also its colour is as a reference of what would be expected: green means it would be expected to have the majority of answers correct, as this part of the test is at the left side of the mark obtained, while red means it would be expected to have the majority of the answers incorrect.

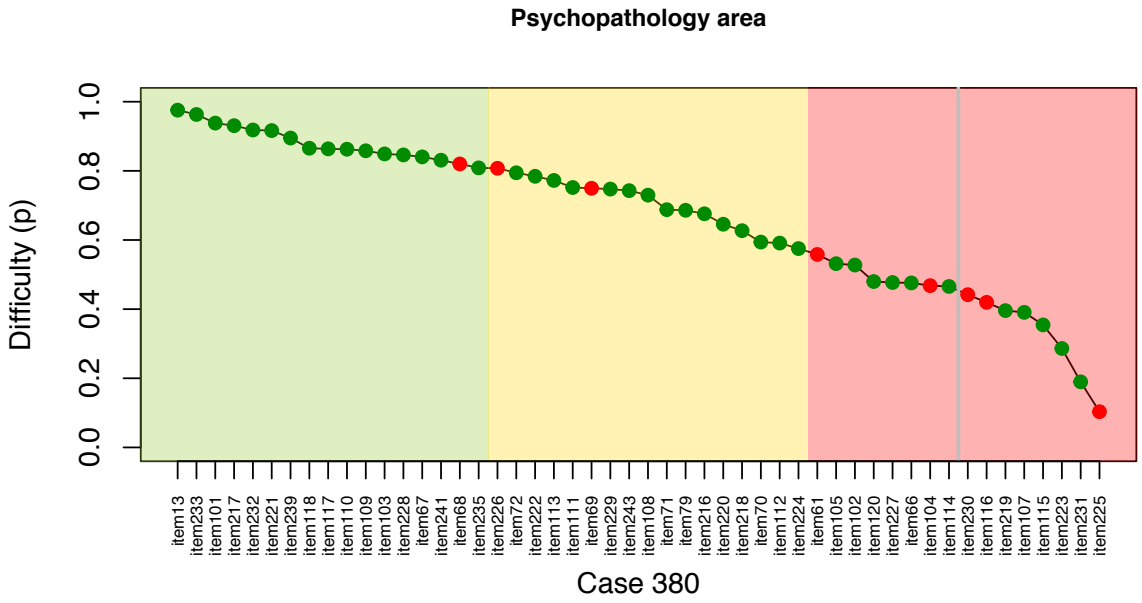


Figure 2: In the x axis the items, ordered by its estimated difficulty are shown, which is reported in the y axis. The vertical line shows the punctuation of the person. Each rectangle of the background is a part of the test, having been this divided into three tertiles of same number of items aproximately: the first area (green) would correspond to the easiest items; the last one (red) would correspond to the most difficult ones and the one in the middle (orange) would correspond to intermediate difficulty items. Green points show an item has been answered correctly and red ones are wrong ansered items. The colors of the background indicate the tertiles in which the items can be divided.

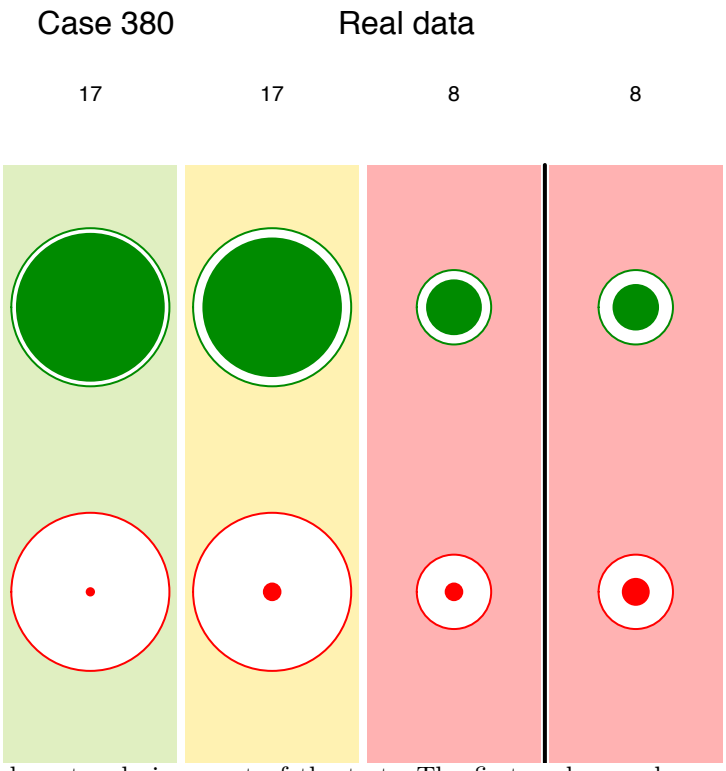


Figure 3: Each rectangle is a part of the test. The first and second ones correspond to the first and second tertiles while the last one corresponds to the third tertile, divided by the mark obtained by the person. In each rectangle the green coloured circles indicate the proportion of correct answers obtained, while the red ones indicate the proportion of wrong answers for that part of the test. The colour of the circle around them indicates what would be expected: green for correct answers and red for incorrect ones.

After having explained these plots, we can proceed to simulate different responses patterns, regarding the person's ability for this case and the item parameters of this section of the test. Responses for cheaters, creative respondents, guessing and careless respondents have been simulated, following the method used in Karabatsos (2003). Cheaters would be answering correctly difficult questions, even though they do not have the level to do so, creative respondents would do the opposite, obtaining incorrect responses to easy items, because of a creative interpretion of them, guessing happens when the examinee guesses the correct answers to some test items, for which he/she does not know the correct answer and careless respondents answer items they know wrongly because they do not pay enough attention. The plots for each of this simulated patterns can be seen in Figure 4.

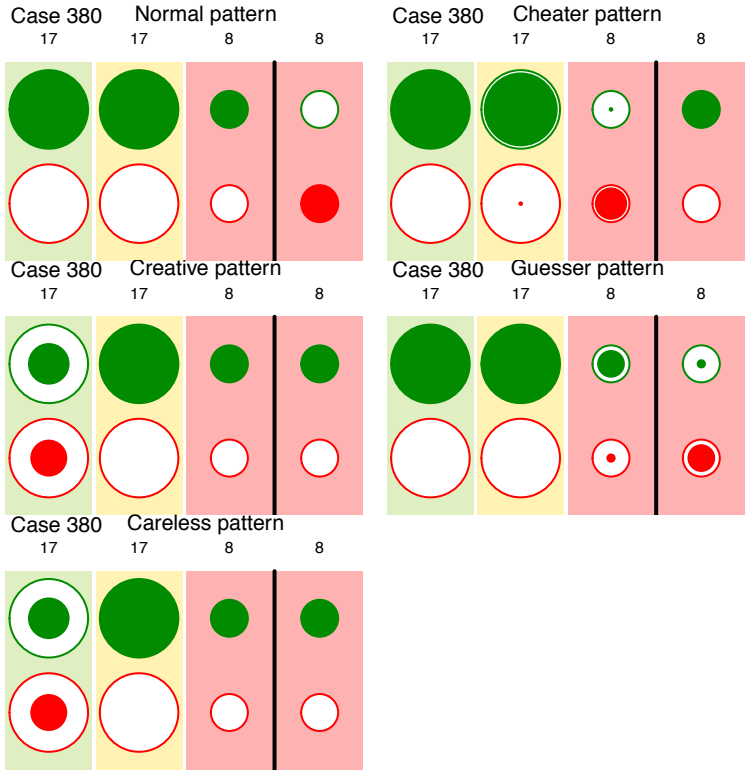


Figure 4: Patterns for the different types of respondants, fixed the level's ability at respondant's 380 level.

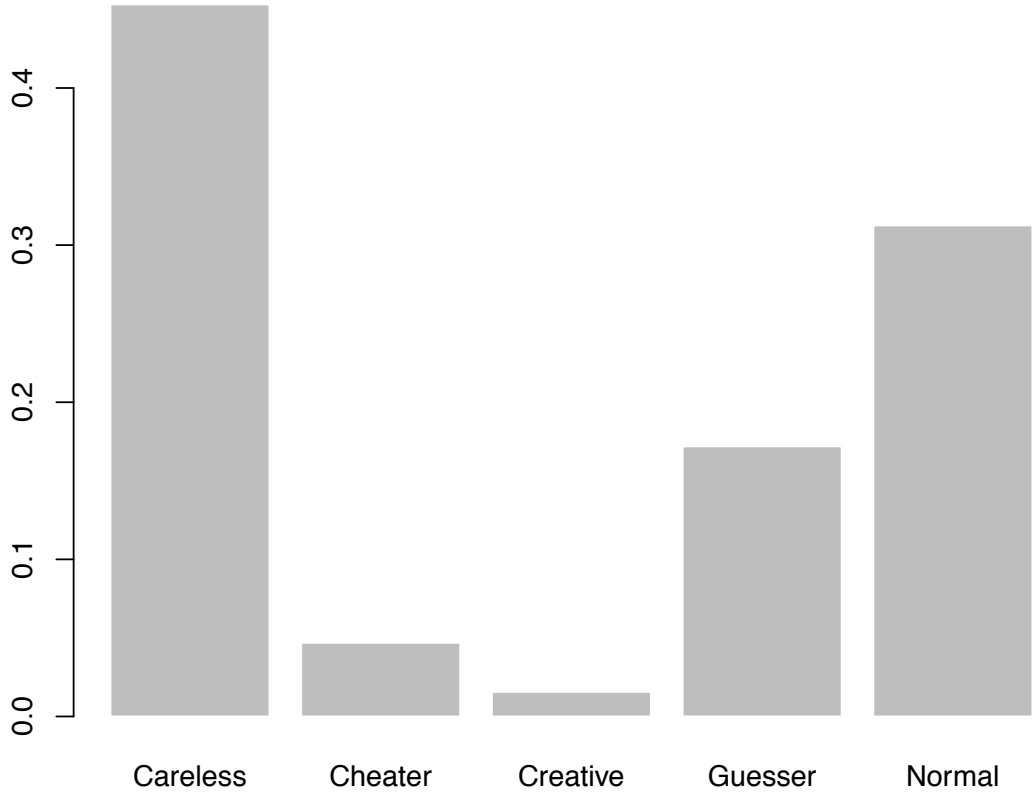
For the case being analysed (case 380), we can see that, by comparing the plots obtained, he/she could actually be classified as a lucky guesser or a careless/creative respondent. As it is difficult and dicussible to discriminate to which profile belongs every person and in order to set a quantitative indicator, the euclidean distance can be computed between the observed pattern on the data and the simulated ones. The distance has been computed between the four rectangles to which the plot is divided (splitting by tertiles and the punctuation). The results obtained can be checked in Table 5 and show that by this criterion this person could be classified either as a creative or a careless respondent, both profiles associated to a higher punctuation than what the mark suggests.

Normal	Cheater	Creative	Lucky Guesser	Careless
8.37	6.48	4.69	6.48	4.69

Table 5: Euclidean distance between the number of correct items for case 380 in the Psychodiagnostic dataset and the simulated profiles for this case. The number of correct items are accounted after splitting the items into groups by the tertiles and the punctuation obtained by the person.

The same strategy has been followed for each of the cases marked as aberrant respondants. Up to the 50% of them were classified as careless respondants, which would mean their score underrated their real knowdledge. About the other 50%, nearly the 30% were classified as normal respondents, not agreeing with the indices performed; around the 15% were classified as guessers, what would imply their real mark overrated their real knowdlege, and the same could be said about the resting 5%, classified as cheaters.

In order to exemplify these results and going back to the plots, let us see some of the cases (Figure 5). Individual 457 in the Psychopathology area was classified as a normal respondent, which does not present concordance with the fact that this individual had been flagged by some of the chosen person-fit indices for this area. Nevertheless, from the plot it can actually be seen that he/she does actually present a normal pattern. Case 426 in the Therapies and treatments area was classified as a lucky guesser, as the plot shows this person answers correctly some of the most difficult questions (above his/her level). Case 558 was classified as a careless respondent, as some of the easiest questions were answered wrongly. Finally, the only case being classified as a cheater would be case 603, as this person answers correctly a higher percentage of the most difficult items than what he/she is expected to. However, of course, it is important to have in mind in every moment that these are just inferences and it would be neeeded further information in order to obtain an accurate conclusion.



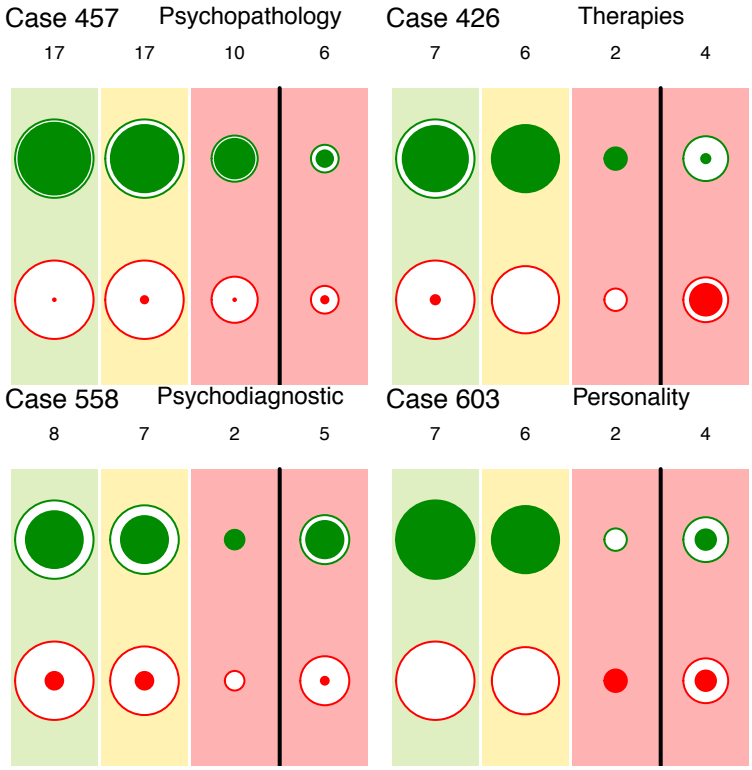


Figure 5: Distribution of the response patterns of some of the respondents that obtained the best punctuations on the test and are flagged as aberrant respondents by the person-fit indices in some of the areas. The individual shown in the top left finally be classified as a normal respondent, the one from the top right as a lucky guesser, the one in the bottom left as a careless respondent and the last one as a cheater.

5 Discussion

Through this project it has been understood the process of performing a person-fit analysis in order to detect aberrant response patterns in an exam's data, and it has been performed on a dataset from the PIR exam of 2005. All the process explained through the project could be considered as a screeing tool in order to identify possible aberrant response patterns in exams, which could imply the inference made from its score is not correct. In this case, as it was one of the objectives of the test, just the people who opted to obtain the internship were analysed, but it could also have been interesting to analyse those who did not get it but were close to it, in order to provide them some advice to improve next time. However, as it has already been said, this process can only be considered as a screening tool, for several reasons: first of all because visual identification of the patterns is not always clear but also because the implications of the accusations of presenting some of the patterns mentioned is not banal. Therefore, it would be needed more information to complement the process, such as interviews with the flagged respondents, in order to take proper conclusions.

Finally, a few limitations of this study are worth mentioning: First of all, for the

person-fit analysis, not answered items have been considered as not known items, imputing the values with 0s. Instead of this, a more sofisticated procedure could have been used, such as multiple imputation. Also, as the data from the exam is not accessible, and also regarding ethical concerns, individuals who had commited aberrant response patterns could not be identified and further explored beyond the dataset. Having greater access to the data could have provided a much more informative analysis, as information of interviews and other resources could have been incorporated. Also, that way it could have been of much more utility, as the results could have been used as feedback to the test-takers, who could have improved its techniques for following convocatoires.

This last concern could lead to future lines: performing this kind of analysis in more accessible groups. For example, teachers could apply person-fit analysis to the results of their exams, as the results from the PIR of 2005 cannot be changed, and now feedback cannot be provided, but it can be done of an exam recently done with properly accessible respondents.

References

[1] Abad, F., Olea, J., Ponsoda, V. & García, C. (2014). *Medición en ciencias sociales y de la salud.* Madrid: Síntesis.

[2] Baker, F. & Kim, S. (2010) *Item Response Theory: Parameter Estimation Techniques.* New York, NY: Dekker.

[3] Conijn, J. M., Emons, W. H., Jong, K. D., & Sijtsma, K. (2015). Detecting and Explaining Aberrant Responding to the Outcome Questionnaire-45. Assessment, 24(4), 513-524. doi:10.1177/1073191114560882

[4] Drasgow, F. and Lissak, R. (1983) Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology*, 68, 363-373.

[5] Drasgow, F., Levine, M. V., and Williams, E. A. (1985) *Appropriateness measurement with poly- chotomous item response models and standardized indices.* British Journal of Mathematical and Statistical Psychology, 38(1), 67âĂŞ86.

[6] Harnisch, D. L., and Linn, R. L. (1981) *Analysis of item response patterns: Questionable test data and dissimilar curriculum practices.* Journal of Educational Measurement, 18(3), 133âĂŞ146.

[7] Junker, B. W., & Sijtsma, K. (2000). *Latent and manifest monotonicity in item response models.* Applied Psychological Measurement, 24, 65âĂŞ81.

[8] Karabatsos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied measurment in education*, 16(4), 277-298.

[9] Krause, M. S. (2017, July 4). Item response theory requires logically unjustifiable assumptions. *Quality & Quantity*, 51, 1549-1561. doi:10.1007/s11135-016-0351-0

[10] Jang, E. E., & Roussos, L. (2007). An investigation into the dimensionality of TOEFL using conditional covariance-based nonparametric approach. *Journal of Educational Measurement*, 44, 1-21.

[11] Meijer, R. R. (1996). Person-Fit Research: An Introduction. *Applied Measurement in Education*, 9(1), 3-8. doi:10.1207/s15324818ame0901_2

[12] Meijer, R. R. (1997, June 2). Person Fit and Criterion-Related Validity: An Extension of the Schmitt, Cortina, and Whitney Study. *Applied Psychological Maasurment*, 21(2), 99-113. doi:10.1177/01466216970212001

[13] Meijer, R. R. , & Sijtsma, K. (1995a). Detection of aberrant item score patterns: A review and new developments. *Applied Measurement in Education*, 8, 261-272. doi:10.1177/01466210122031957

[14] Meijer, R. R., & Sijtsma, K. (1995b). Person-fit analysis: Classification of persons on the basis of their item score patterns. In J. J. Hox, & W. Jansen (Eds.), *Measurement problems in social and behavioral research* (pp. 51-66). (SCO-rapport; No. 381). Amsterdam: Stichting Kohnstamm Fonds voor Onderwijsresearch.

[15] Mokken, R. J. (1971) *A theory and procedure of scale analysis.* Berlin, Germany: De Gruyter.

[16] Molenaar IW, Sijtsma K (2000). User's Manual **MSP5** for Windows [computer software and manual]. IEC ProGAMMA, Groningen, The Netherlands.

[17] Núñez, R. M., & López, J. A. (2006, June). Técnicas para detectar patrones de respuesta atípicos. *Anales de Psicología*, 22(1), 143-154. URL http://revistas.um.es/analesps/

[18] Rizopoulos, D. (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, 17 (5), 1-25. URL http://www.jstatsoft.org/v17/i05/

[19] Sato, T. (1975) *The construction and interpretation of S-P tables.* Tokyo: Meiji Tosho.

[20] Sijtsma, K. (1986) *A coefficient of deviance of response patterns.* Kwantitatieve Methoden, 7, 131-145.

[21] Sijtsma, K, and Molenaar, I. W. (2002) *Introduction to nonparametric item response theory.* Thousand Oaks, CA: Sage.

[22] Tendeiro, J. N., Meijer, R. R., & Niessen, A. M. (2016, October). PerFit: An R Package for Person-Fit Analysis in IRT. *Journal of Statistical Software*, 74(5). doi:10.18637/jss.v074.i05

[23] van der Flier, H. (1982) *Deviant response patterns and comparability of test scores.* Journal of Cross-Cultural Psychology, 13(3), 267âĂŞ298.

A Descriptive analysis plots

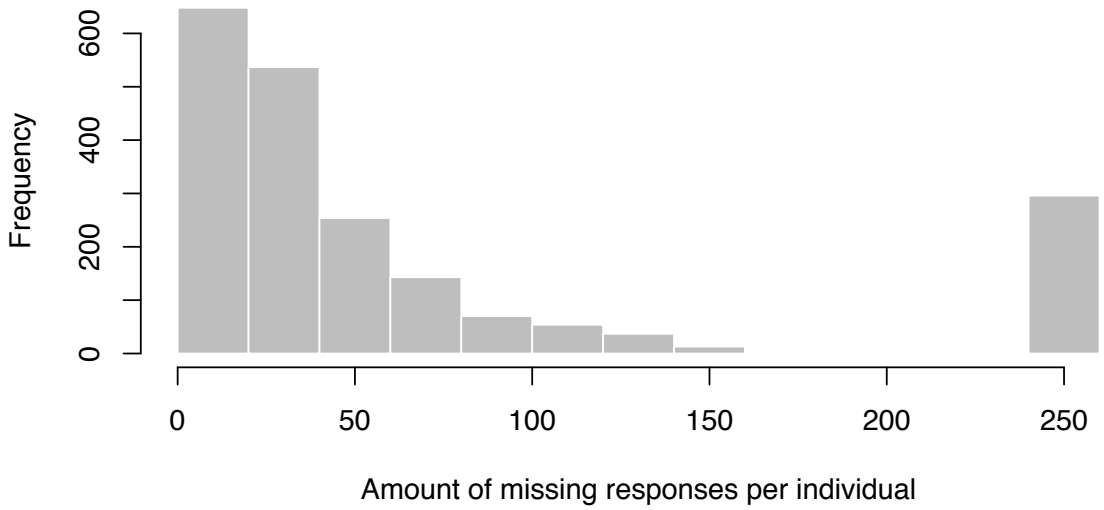


Figure 6: Histogram of the number of missing responses per case.

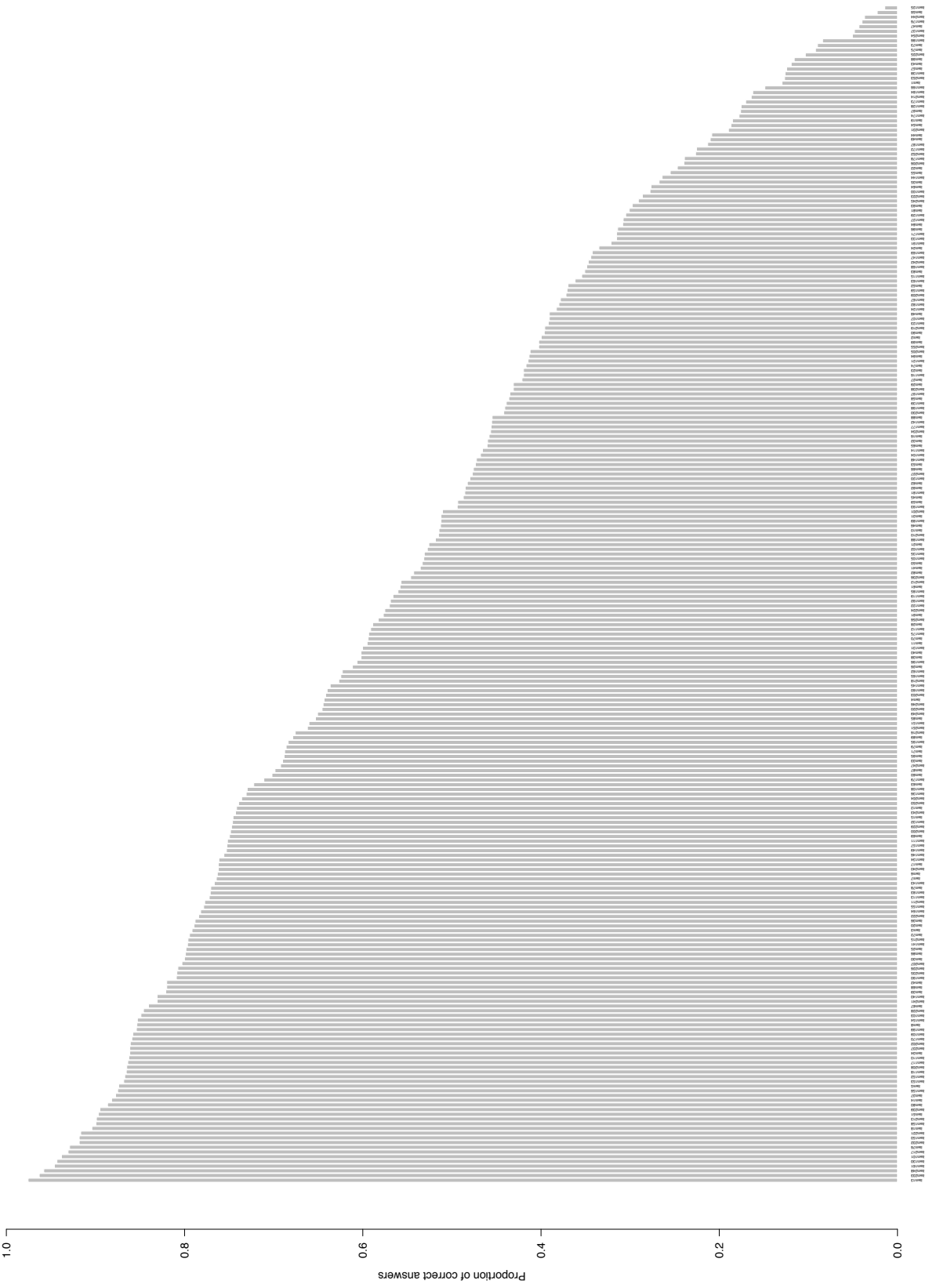


Figure 7: Decreasing proportions of correct answers on all of the items of the test.

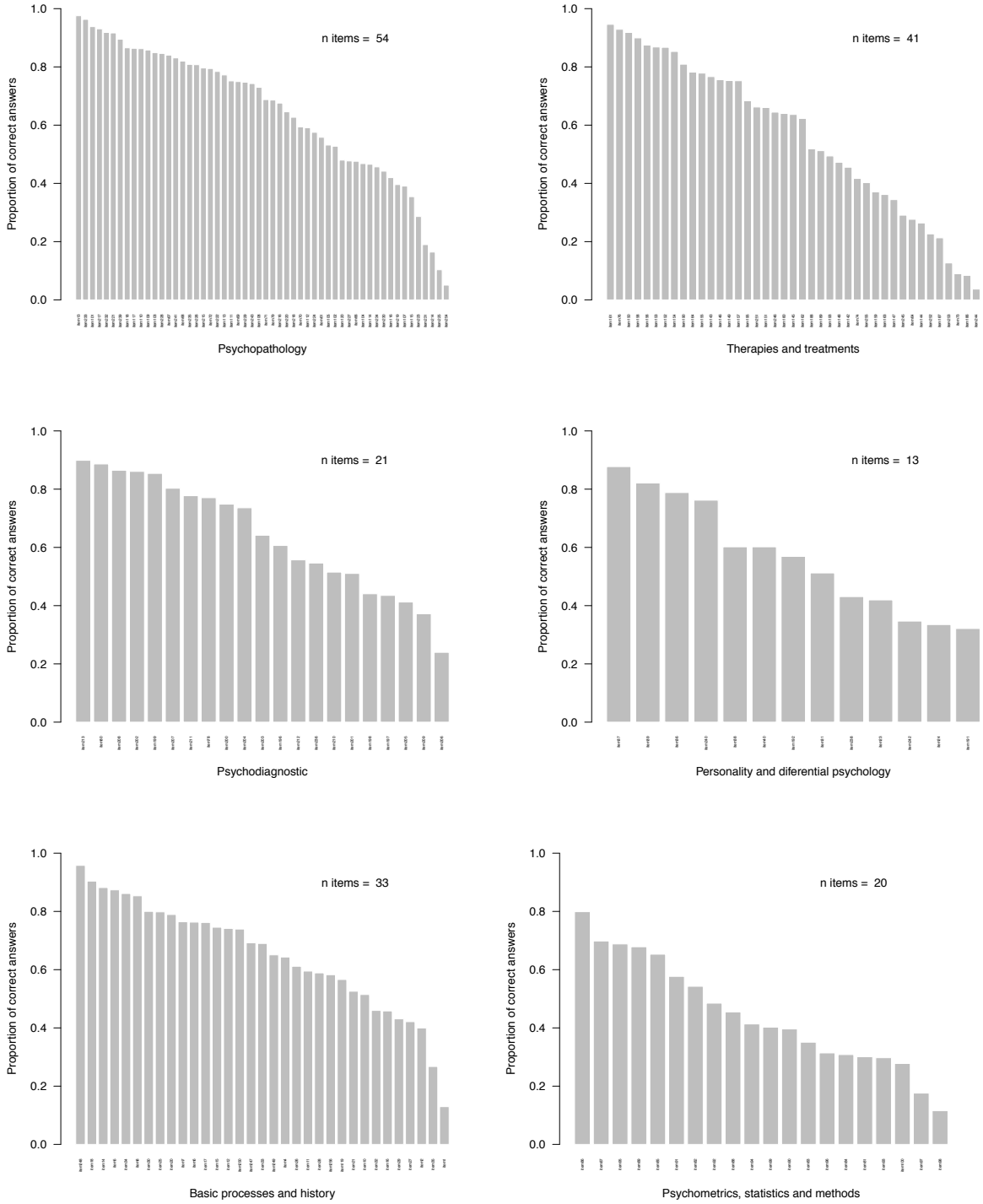


Figure 8: Decreasing proportions of correct answers on the items of the test divided by areas of knowledge or subtests.

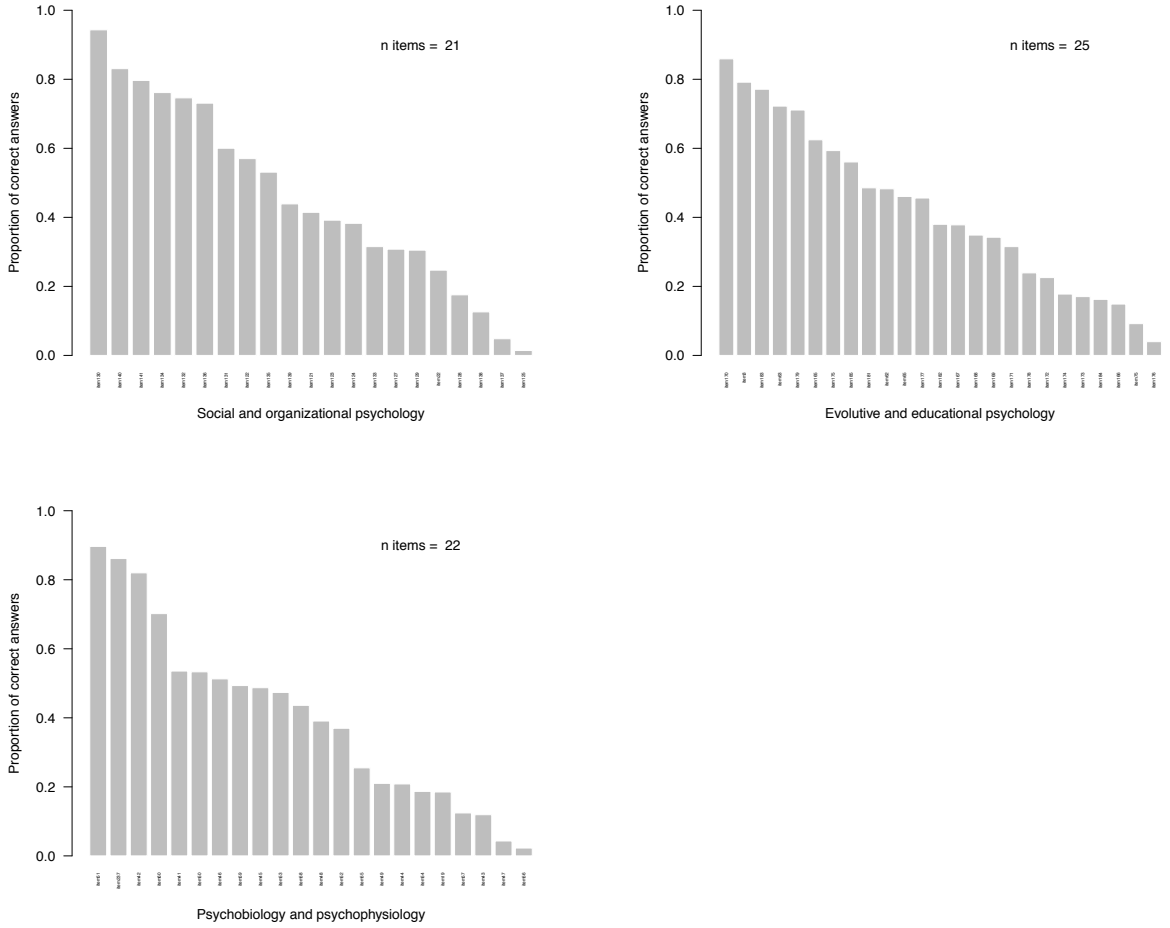


Figure 9: Decreasing proportions of correct answers on the items of the test divided by areas of knowledge or subtests.

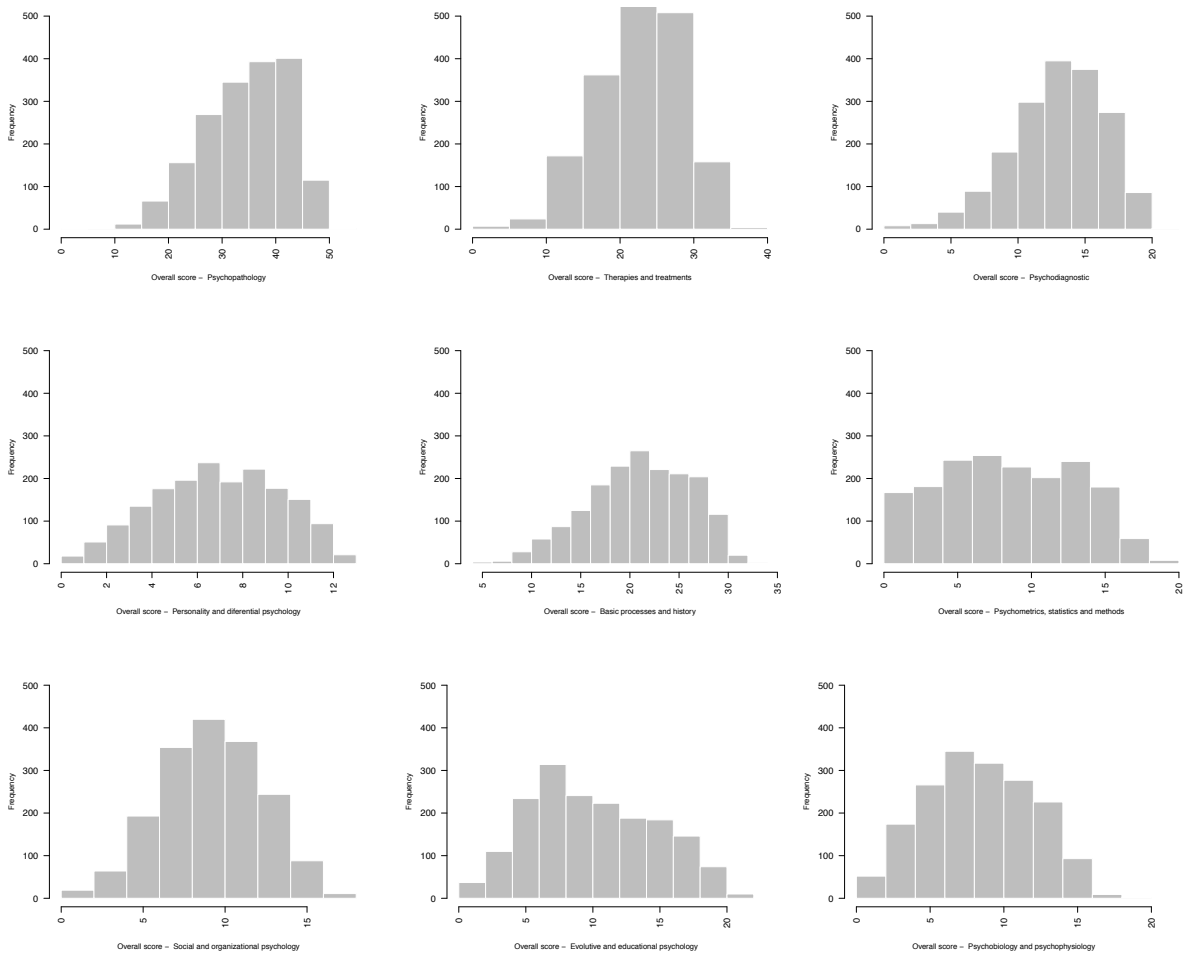


Figure 10: Distribution of the number of correct responses per person by areas of knowledge of the exam.

B Assumption of monotonicity

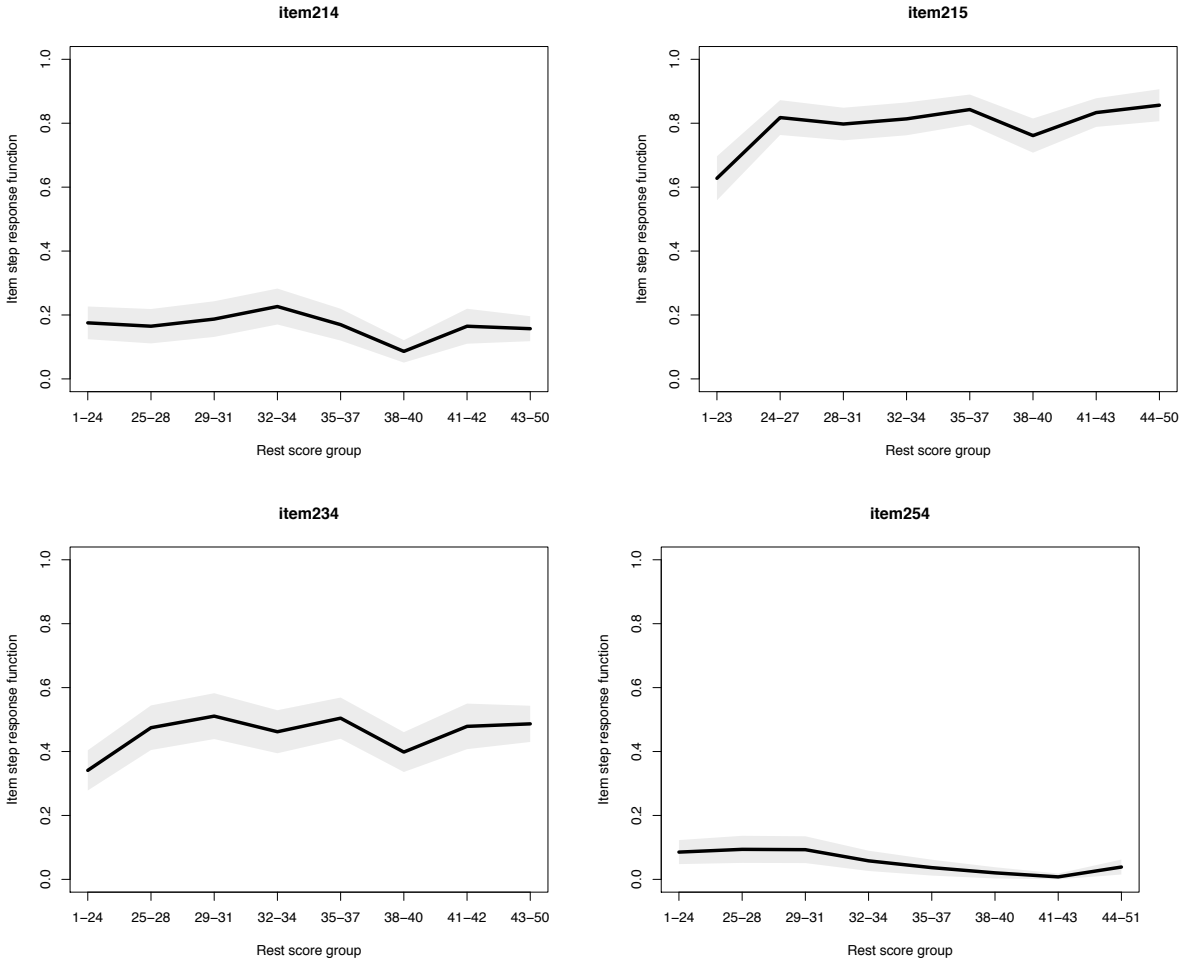


Figure 11: Estimated item step response functions of items with significant violations of manifest monotonicity for the psychopathology subtest items.

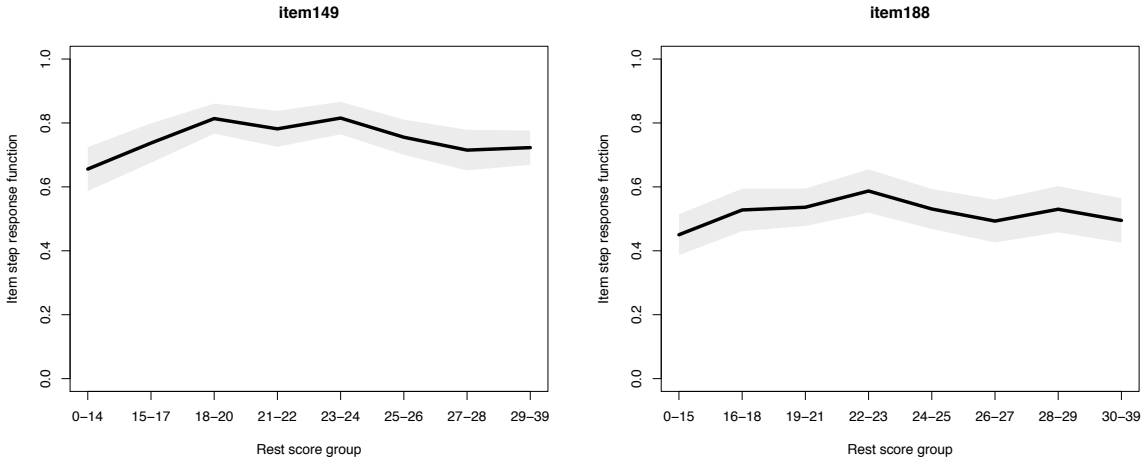


Figure 12: Estimated item step response functions of items with significant violations of manifest monotonicity for the therapies and treatments subtest items.

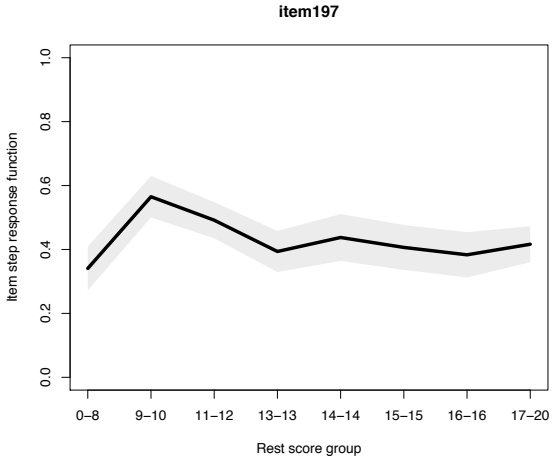


Figure 13: Estimated item step response functions of items with significant violations of manifest monotonicity for the psychodiagnostic subtest items.

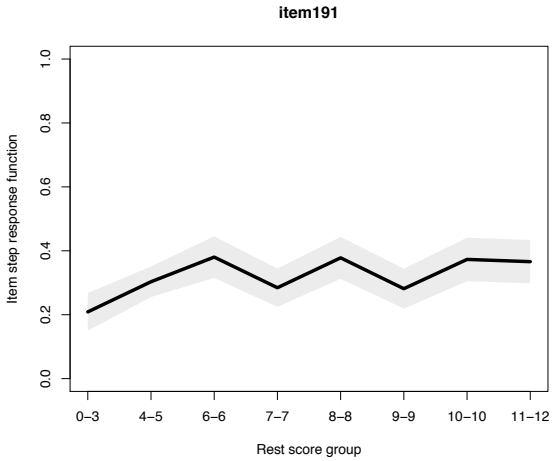


Figure 14: Estimated item step response functions of items with significant violations of manifest monotonicity for the personality and diferential psychology subtest items.

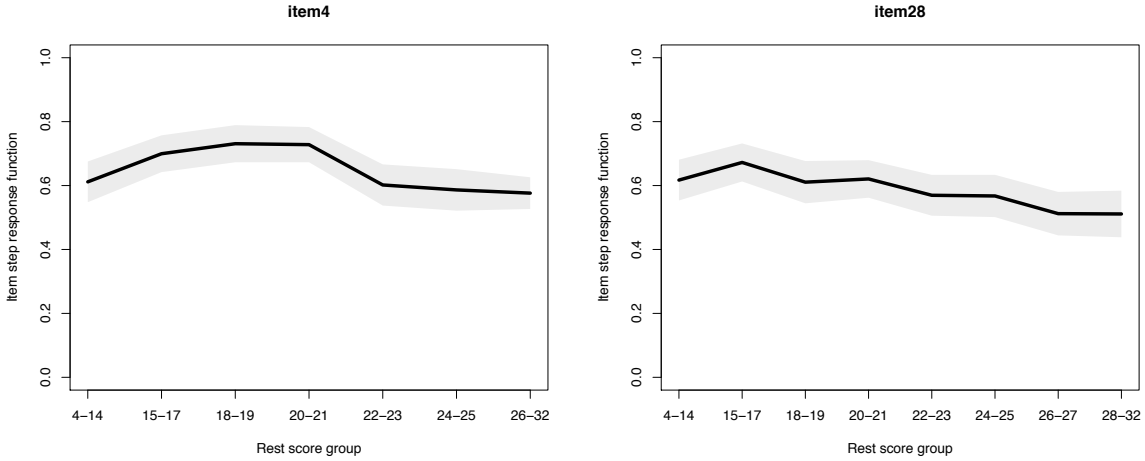


Figure 15: Estimated item step response functions of items with significant violations of manifest monotonicity for the basic processes and history subtest items.

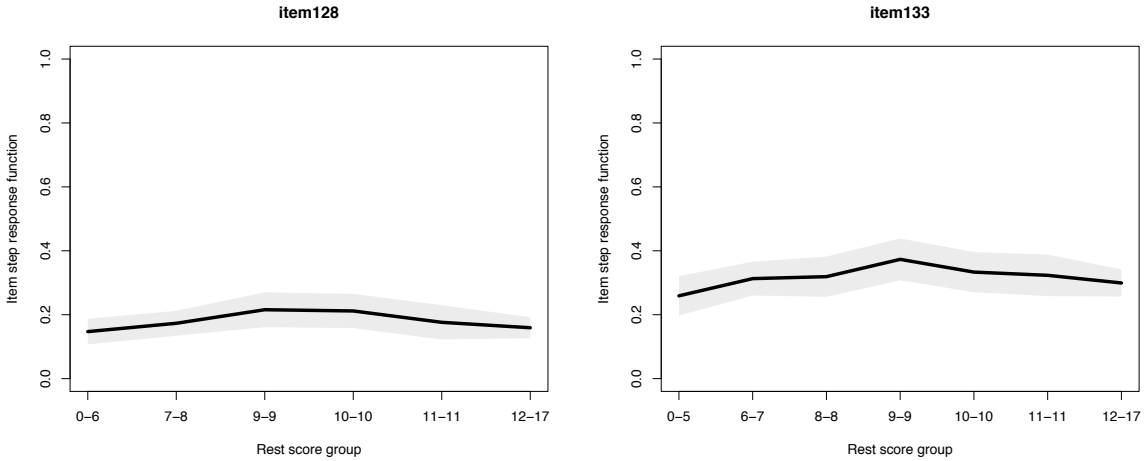


Figure 16: Estimated item step response functions of items with significant violations of manifest monotonicity for the social and organizational psychology subtest items.

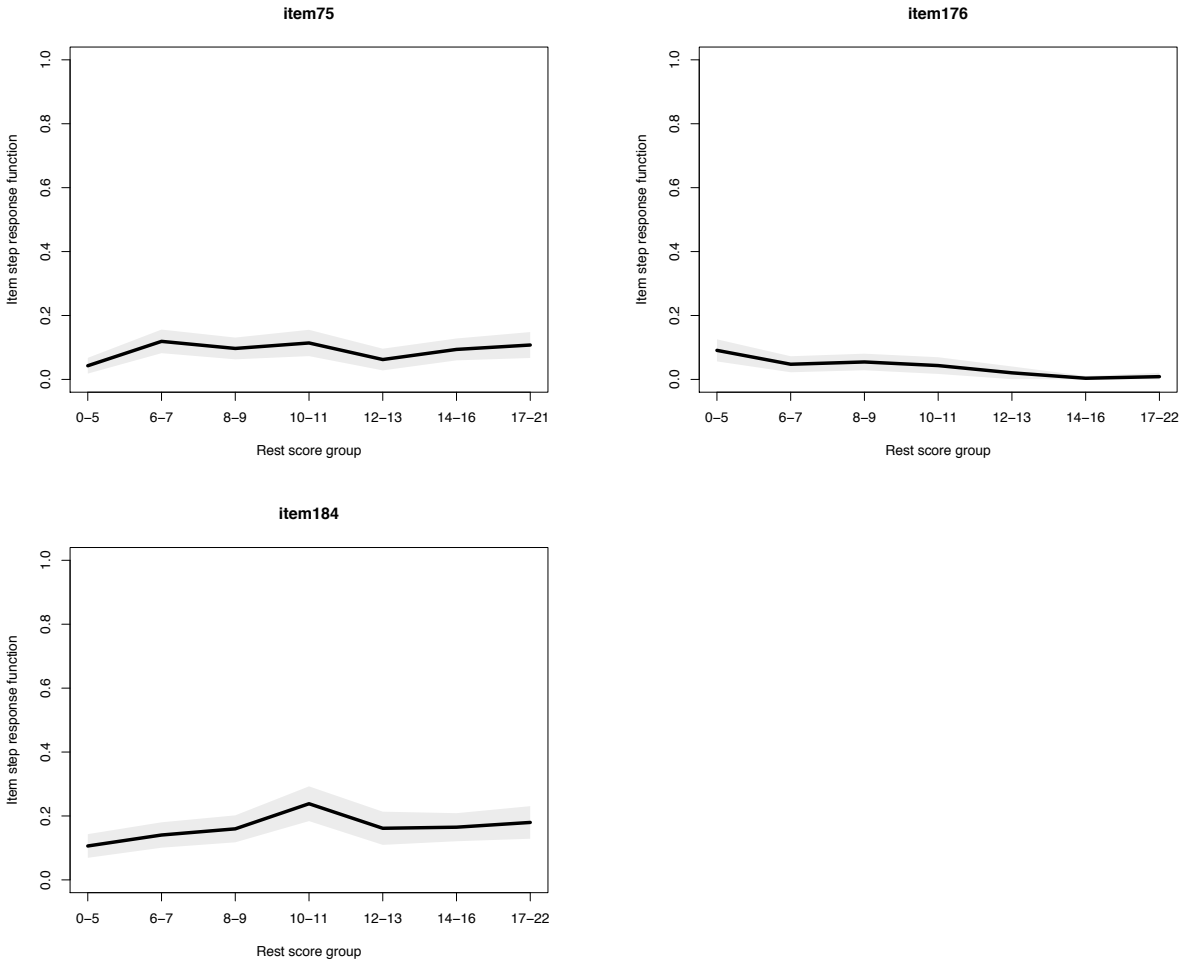


Figure 17: Estimated item step response functions of items with significant violations of manifest monotonicity for the evolutive and educational psychology subtest items.

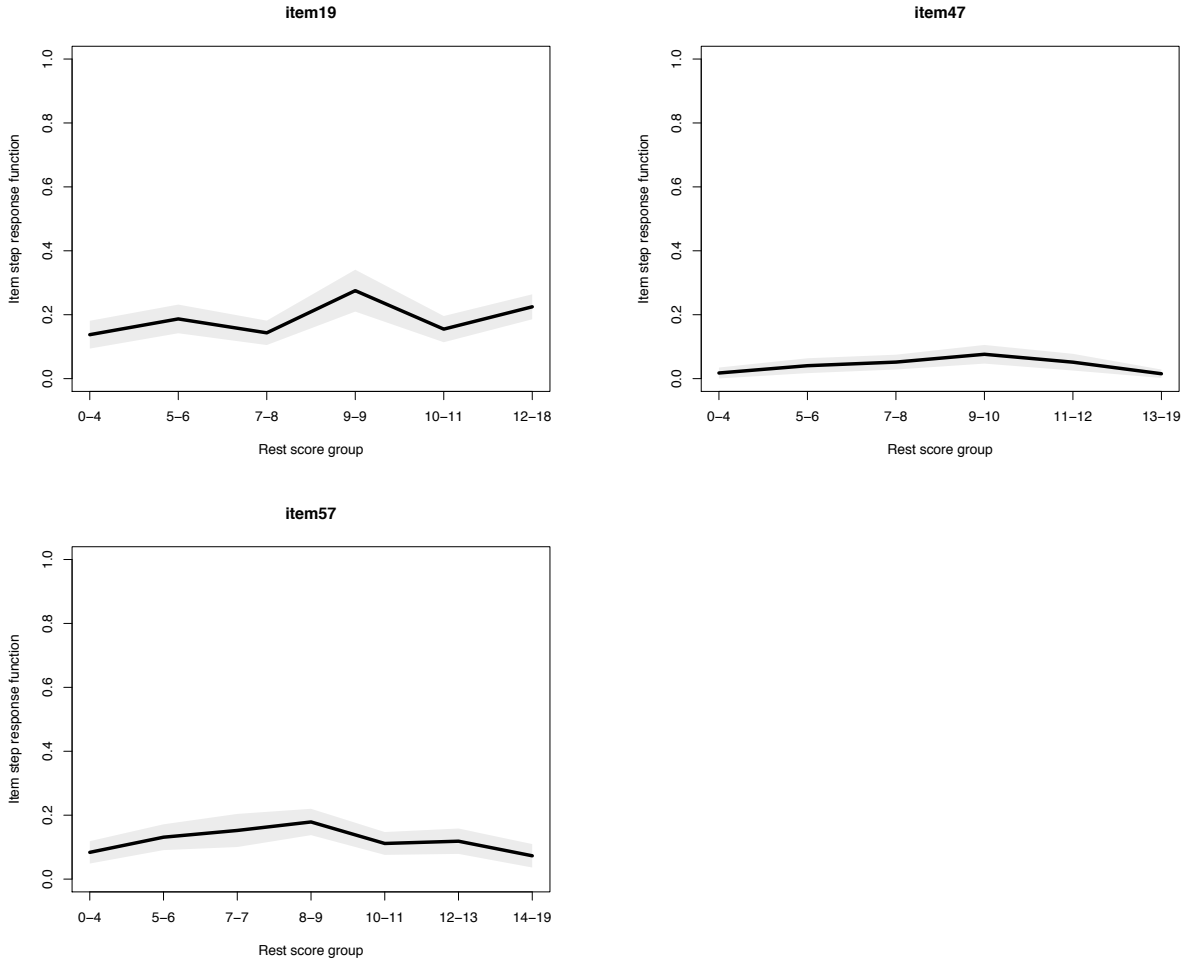


Figure 18: Estimated item step response functions of items with significant violations of manifest monotonicity for the psychobiology and psychophysiology subtest items.

C Assumption of local independency and goodness of fit

Therapies and treatments

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	39	0	0	0	0	0	0	0.00	0.00
Doublets	341	72	57	29	29	31	182	5.03	9.31
Triplets	1656	1101	873	715	604	1090	3100	6.53	7.56

Table 6: Results from the analysis of the local independence and the fit of the 1PL IRT model for the Therapies and treatments data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	39	0	0	0	0	0	0	0.00	0.00
Doublets	608	72	24	13	12	6	6	0.56	1.27
Triplets	6107	1702	752	313	138	101	26	0.92	1.20

Table 7: Results from the analysis of the local independence and the fit of the 2PL IRT model for the Therapies and treatments data.

Psychodiagnostic

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	20	0	0	0	0	0	0	0.00	0.00
Doublets	97	21	6	9	15	11	31	3.64	6.51
Triplets	225	166	151	101	88	138	271	4.76	4.86

Table 8: Results from the analysis of the local independence and the fit of the 1PL IRT model for the Psychodiagnostic data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	20	0	0	0	0	0	0	0.00	0.00
Doublets	162	13	12	2	1	0	0	0.42	0.81
Triplets	790	251	76	17	6	0	0	0.76	0.82

Table 9: Results from the analysis of the local independence and the fit of the 2PL IRT model for the Psychodiagnostic data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	20	0	0	0	0	0	0	0.00	0.00
Doublets	163	12	11	2	2	0	0	0.41	0.82
Triplets	792	254	74	16	3	1	0	0.74	0.81

Table 10: Results from the analysis of the local independence and the fit of the 3PL IRT model for the Psychodiagnostic data.

Personality and diferential psychology

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	12	0	0	0	0	0	0	0.00	0.00
Doublets	24	3	5	2	5	4	23	6.38	8.02
Triplets	26	8	24	13	15	21	113	8.27	6.33

Table 11: Results from the analysis of the local independence and the fit of the 1PL IRT model for the Personality and diferential psychology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	12	0	0	0	0	0	0	0.00	0.00
Doublets	62	3	1	0	0	0	0	0.18	0.47
Triplets	187	28	4	1	0	0	0	0.42	0.60

Table 12: Results from the analysis of the local independence and the fit of the 2PL IRT model for the Personality and diferential psychology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	12	0	0	0	0	0	0	0.00	0.00
Doublets	63	3	0	0	0	0	0	0.15	0.41
Triplets	205	13	2	0	0	0	0	0.27	0.43

Table 13: Results from the analysis of the local independence and the fit of the 3PL IRT model for the Personality and diferential psychology data.

Basic processes and history

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	31	0	0	0	0	0	0	0.00	0.00
Doublets	191	50	23	24	37	40	100	5.08	9.69
Triplets	572	419	570	488	404	621	1421	6.57	7.68

Table 14: Results from the analysis of the local independence and the fit of the 1PL IRT model for the Basic processes and history data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	31	0	0	0	0	0	0	0.02	0.08
Doublets	403	32	14	8	3	3	2	0.42	1.06
Triplets	3286	758	275	101	37	34	4	0.73	1.01

Table 15: Results from the analysis of the local independence and the fit of the 2PL IRT model for the Basic processes and history data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	31	0	0	0	0	0	0	0.00	0.00
Doublets	411	29	9	9	2	5	0	0.36	0.93
Triplets	3554	583	248	75	21	13	1	0.58	0.86

Table 16: Results from the analysis of the local independence and the fit of the 3PL IRT model for the Basic processes and history data.

Psychometrics, statistics and methods

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	20	0	0	0	0	0	0	0.00	0.00
Doublets	50	17	11	12	3	13	84	10.79	15.59
Triplets	26	50	80	60	56	118	750	13.41	12.23

Table 17: Results from the analysis of the local independence and the fit of the 1PL IRT model for the Psychometrics, statistics and methods data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	20	0	0	0	0	0	0	0.00	0.00
Doublets	153	19	11	3	0	1	3	0.84	3.25
Triplets	664	260	103	53	14	10	36	1.47	2.70

Table 18: Results from the analysis of the local independence and the fit of the 2PL IRT model for the Psychometrics, statistics and methods data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	20	0	0	0	0	0	0	0.00	0.00
Doublets	156	16	11	2	1	1	3	0.79	3.00
Triplets	665	265	106	46	16	6	36	1.40	2.53

Table 19: Results from the analysis of the local independence and the fit of the 3PL IRT model for the Psychometrics, statistics and methods data.

Social and organizational psychology

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	19	0	0	0	0	0	0	0.00	0.00
Doublets	79	14	13	14	10	13	28	4.62	10.55
Triplets	147	139	142	113	86	120	222	6.25	8.68

Table 20: Results from the analysis of the local independence and the fit of the 1PL IRT model for the Social and organizational psychology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	19	0	0	0	0	0	0	0.00	0.00
Doublets	136	16	7	6	2	4	0	0.63	1.31
Triplets	612	186	88	40	20	16	7	1.06	1.42

Table 21: Results from the analysis of the local independence and the fit of the 2PL IRT model for the Social and organizational psychology data.

Evolutive and educational psychology

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	22	0	0	0	0	0	0	0.00	0.00
Doublets	69	16	13	6	7	14	106	10.22	11.73
Triplets	98	96	81	87	77	122	979	12.93	10.41

Table 22: Results from the analysis of the local independence and the fit of the 1PLIRT model for the Evolutive and educational psychology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	22	0	0	0	0	0	0	0.01	0.04
Doublets	187	21	12	3	1	7	0	0.57	1.24
Triplets	916	311	190	78	31	14	0	1.09	1.18

Table 23: Results from the analysis of the local independence and the fit of the 2PLIRT model for the Evolutive and educational psychology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	22	0	0	0	0	0	0	0.00	0.00
Doublets	200	11	12	6	1	1	0	0.37	0.91
Triplets	1115	295	88	33	7	2	0	0.68	0.88

Table 24: Results from the analysis of the local independence and the fit of the 3PLIRT model for the Evolutive and educational psychology data.

Psychobiology and psychophysiology

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	19	0	0	0	0	0	0	0.00	0.00
Doublets	79	15	14	6	8	12	37	4.69	8.09
Triplets	131	123	126	107	70	103	309	6.15	6.31

Table 25: Results from the analysis of the local independence and the fit of the 1PLIRT model for the Psychobiology and psychophysiology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	19	0	0	0	0	0	0	0.00	0.00
Doublets	151	10	6	2	2	0	0	0.33	0.84
Triplets	715	170	60	18	5	1	0	0.66	0.88

Table 26: Results from the analysis of the local independence and the fit of the 2PLIRT model for the Psychobiology and psychophysiology data.

	Less 1	1 to 2	2 to 3	3 to 4	4 to 5	5 to 7	Larger 7	Mean	SD
Singlets	19	0	0	0	0	0	0	0.00	0.00
Doublets	153	10	4	2	2	0	0	0.28	0.78
Triplets	767	135	44	17	6	0	0	0.55	0.83

Table 27: Results from the analysis of the local independence and the fit of the 3PLIRT model for the Psychobiology and psychophysiology data.

D Assumption of unidimensionality

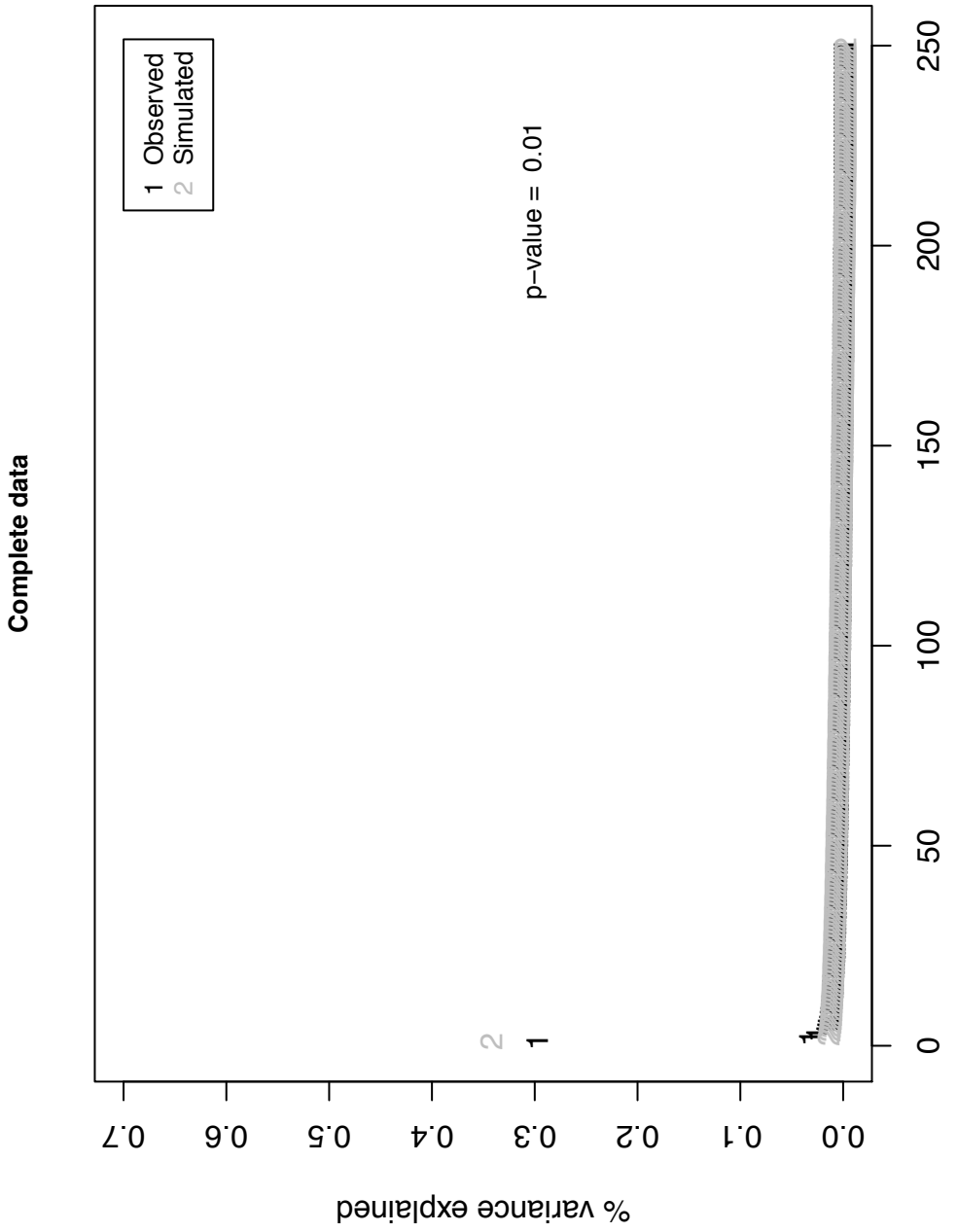


Figure 19: Plot for Unidimensionality Check using Modified Parallel Analysis for the complete dataset. 1s (in black) represent the values for the observed values of the eigenvalues while 2s (in gray) are the mean of the simulated eigenvalues under the assumption of unidimensionality.

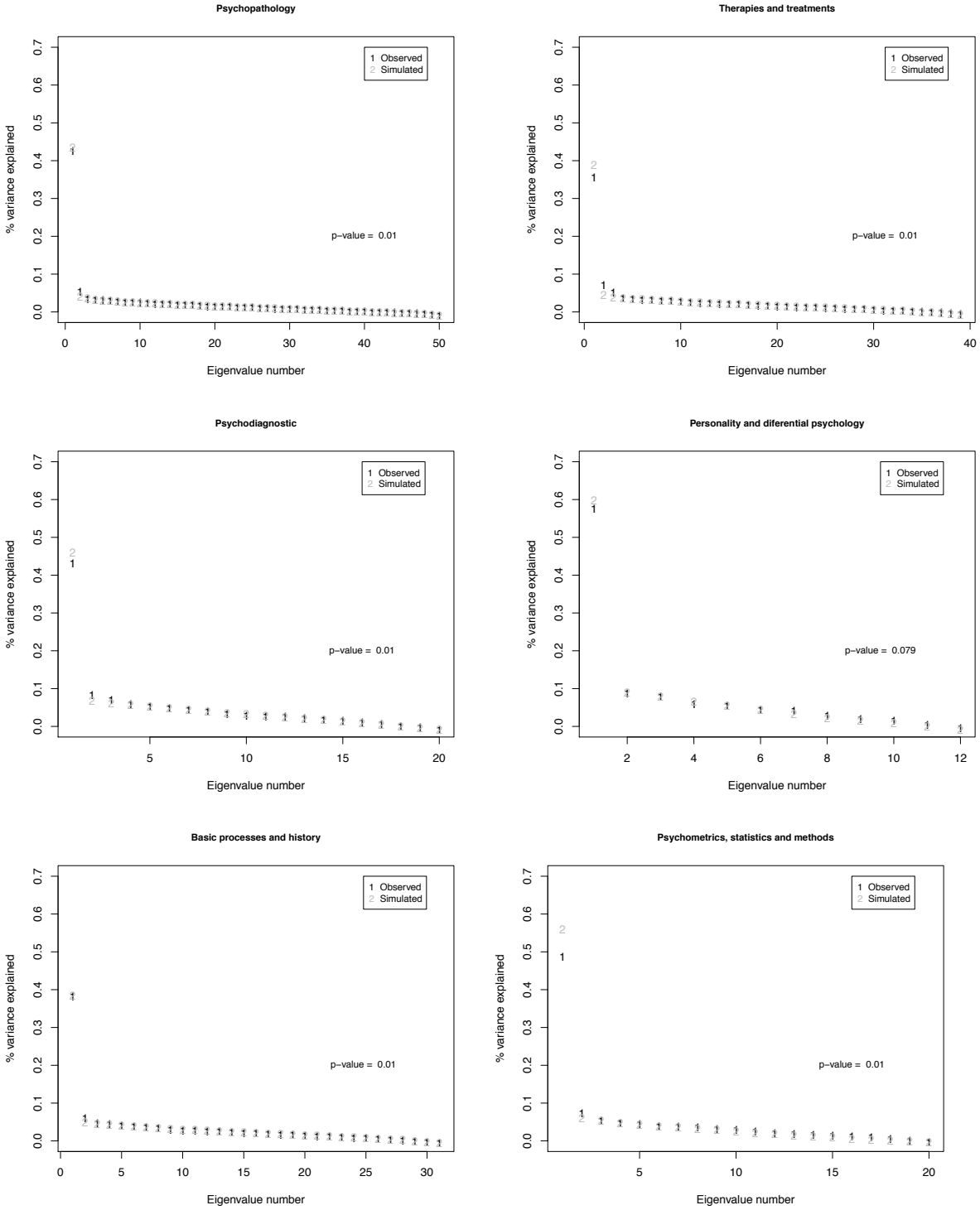


Figure 20: Plots for Unidimensionality Check using Modified Parallel Analysis for the dataset divided by areas of psychology.

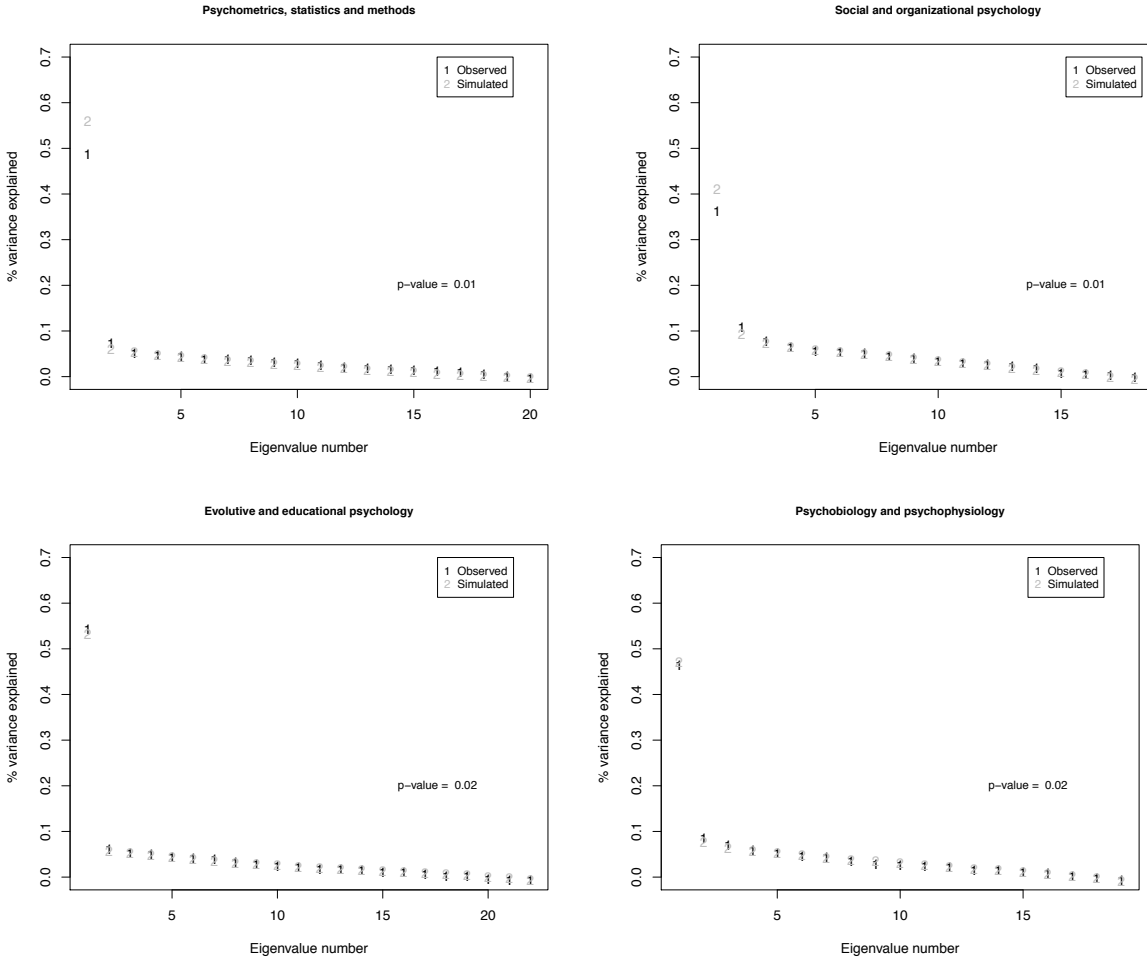


Figure 21: Plots for Unidimensionality Check using Modified Parallel Analysis for the dataset divided by areas of psychology.

E Indices profiles

$C*$	H_t	lz	%	#flags	overall %
●	●	●	2.61	3	2.61
●	●	○	2.44	2	3.92
●	○	●	0.00		
○	●	●	1.48		
●	○	○	0.62	1	4.88
○	●	○	1.14		
○	○	●	3.12		
○	○	○	88.59	0	88.59

Table 28: All the possible profiles for the Psychopathology dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

$C*$	H_t	lz	%	#flags	overall %
●	●	●	4.54	3	4.54
●	●	○	0.85	2	2.04
●	○	●	0.17		
○	●	●	1.02		
●	○	○	0.06	1	2.61
○	●	○	0.91		
○	○	●	1.65		
○	○	○	90.80	0	90.80

Table 29: All the possible profiles for the Therapies and treatments dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

$C*$	H_t	lz	%	#flags	overall %
●	●	●	2.44	3	2.44
●	●	○	1.87	2	2.84
●	○	●	0.00		
○	●	●	0.97		
●	○	○	0.34	1	3.46
○	●	○	1.36		
○	○	●	1.76		
○	○	○	91.25	0	91.25

Table 30: All the possible profiles for the Psychodiagnostic dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

$C*$	H_t	lz	%	#flags	overall %
●	●	●	3.29	3	3.29
●	●	○	2.50	2	2.95
●	○	●	0.45		
○	●	●	0.00		
●	○	○	0.23	1	2.50
○	●	○	0.00		
○	○	●	2.27		
○	○	○	91.25	0	91.25

Table 31: All the possible profiles for the Personality and diferential psychology dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

$C*$	H_t	lz	%	#flags	overall %
●	●	●	3.52	3	3.52
●	●	○	2.50	2	3.52
●	○	●	0.00		
○	●	●	1.02		
●	○	○	0.68	1	4.32
○	●	○	0.68		
○	○	●	2.95		
○	○	○	88.64	0	88.64

Table 32: All the possible profiles for the Basic processes and history dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

$C*$	H_t	lz	%	#flags	overall %
●	●	●	3.24	3	3.24
●	●	○	3.52	2	4.09
●	○	●	0.00		
○	●	●	0.57		
●	○	○	0.34	1	3.35
○	●	○	0.06		
○	○	●	2.95		
○	○	○	89.32	0	89.32

Table 33: All the possible profiles for the Psychometrics, statistics and methods dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

$C*$	H_t	lz	%	#flags	overall %
●	●	●	3.35	3	3.35
●	●	○	2.90	2	3.12
●	○	●	0.23		
○	●	●	0.00		
●	○	○	1.31	1	1.99
○	●	○	0.00		
○	○	●	0.68		
○	○	○	91.54	0	91.54

Table 34: All the possible profiles for the Social and organizational psychology dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

$C*$	H_t	lz	%	#flags	overall %
●	●	●	3.41	3	3.41
●	●	○	1.87	2	2.21
●	○	●	0.11		
○	●	●	0.23		
●	○	○	0.57	1	3.52
○	●	○	0.23		
○	○	●	2.73		
○	○	○	90.86	0	90.86

Table 35: All the possible profiles for the Evolutive and educational psychology dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

$C*$	H_t	lz	%	#flags	overall %
●	●	●	3.41	3	3.41
●	●	○	1.42	2	1.76
●	○	●	0.34		
○	●	●	0.00		
●	○	○	0.91	1	3.29
○	●	○	0.11		
○	○	●	2.27		
○	○	○	91.54	0	91.54

Table 36: All the possible profiles for the Psychobiology and psychophysiology dataset, depending on the configurations of indices. The painted circle indicates presence of the index while the white one means absence of it.

Patrons atípics de resposta en proves de coneixement

Analitzant l'examen PIR de 2005

Els resultats dels tests no sempre reflecteixen els coneixements que es volen avaluar, ja que les respostes poden dependre d'aspectes aliens a aquests. Per exemple, copiar pot inflar la nota, mentre que no posar atenció a les qüestions més fàcils pot fer-la baixar. És important detectar quan les puntuacions estan esbiaixades i no indiquen el nivell real dels coneixements avaluats. Per a fer-ho, es pot dur a terme una anàlisi de patrons atípics de resposta (PAR), que consisteix a identificar respostes de persones concretes que no segueixen una pauta esperada. Es pot realitzar mitjançant un gran número d'índexs. En aquest treball se n'han fet servir tres (H_T , U_3 i l_z) per tal de descobrir si alguna de les 89 persones amb puntuació candidata a una plaça PIR a l'any 2005, d'entre els 2057 presentats, havien presentat un PAR.

Els patrons atípics detectats s'han avaluat amb un procediment gràfic per a identificar-ne el tipus de biaix associat.

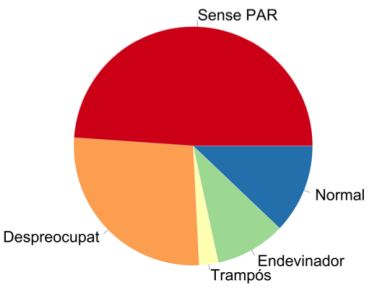
A la figura s'hi mostra el perfil de respostes d'una persona a les 22 preguntes de continguts de psicologia evolutiva, dividides en tres nivells de dificultat (fàcil en verd, mitjà en taronja, difícil en vermell). L'últim nivell es troba dividit en dos blocs ja que la persona ha respost correctament 18 ítems (8 + 7 + 3), pel que abans de la línea negra s'espera respostes correctes (rectangle exterior verd) i després se n'espera d'incorrectes (rectangle exterior vermell). En cada nivell els cercles són proporcionals al seu nombre d'ítems i el color interior n'indica la proporció de respostes correctes (verd) o incorrectes (vermell).

A la figura hi veiem el gràfic que representa el patró observat en la persona, el patró esperat o normal i el resultat de les simulacions de 4 perfils de resposta atípica. Veiem que el perfil presentat s'assembla al d'una persona *creativa* o *descuidada*, ja que ha contestat erròniament a preguntes fàcils de l'examen i ha respost bé algunes de les més difícils, pel que de no haver comès un PAR la seva puntuació podria haver estat més alta. En altres casos detectats s'ha donat el cas contrari, el que fa posar en dubte el nivell de coneixements acreditats en aquesta matèria.

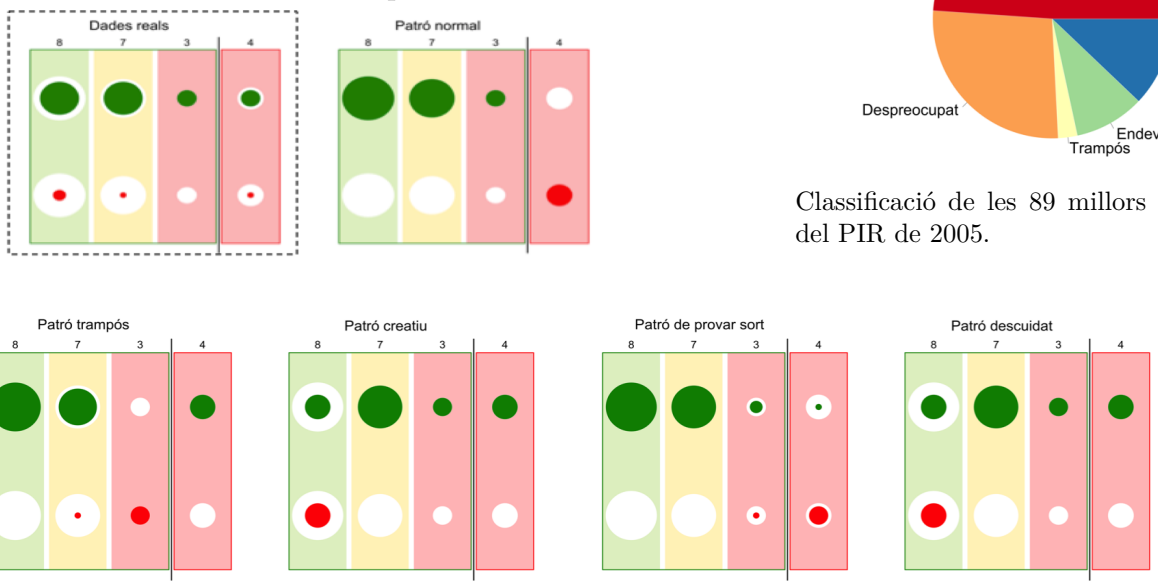
Alguns dels resultats obtinguts a l'anàlisi de la prova PIR de 2005 indiquen que:

- Aproximadament la meitat de les 89 millors puntuacions s'han obtingut de manera atípica.
- D'aquells que presentaven PAR, el 50% tenien un perfil comparable a persones que descuidaven les respostes a les preguntes més fàcils. En aquests casos, probablement la puntuació obtinguda infravalorava els seus coneixements.
- Al voltant del 15% dels PAR eren comparables a persones que, sense tenir un nivell molt alt, proven d'endevinar les preguntes més difícils de la prova.
- En un dels casos (5%), el patró atípic identificat podria correspondre a una persona que ha fet trampes a l'examen.
- La resta dels casos identificats com a PAR no s'allunyaven gaire d'un perfil de respostes normal.

Alba Morató Catafal
Facultat de Psicologia, UAB
albamrt11@gmail.com



Classificació de les 89 millors puntuacions del PIR de 2005.



Perfil de respostes d'una de les persones que han fet l'examen PIR, perfil de respostes normals i perfil de respostes atípiques simulades a partir de la puntuació a la prova.

Resum executiu
Anàlisi de patrons atípics de resposta en tests educatius de resposta múltiple

Alumna: Alba Morató Catafal
Supervisor: Eduardo Doval Diéguez

31 de maig de 2018

El present estudi s'enmarca en l'àmbit de l'avaluació dels coneixements i va dirigit a tots els agents i institucions que hi tenen implicació.
L'objectiu principal és obtenir evidencia de la validesa de les inferències sobre els coneixements de cada persona avaluada, fets a partir de la puntuació obtinguda a la prova. Alguns factors externs al propi coneixement avaluat poden fer que aquest quedi sobre o subestimat. En un context d'avaluació sumativa o certificadora, la identificació individual de la presència d'aquest biaix hauria de ser d'interès per a la persona avaluada, a qui se l'ha de garantir una avaluació vàlida, i també a les persones i institucions encarregades de realitzar-la, ja que són les garants d'aquest dret. En un context d'avaluació diagnòstica o formativa, la identificació d'aquests factors ha de permetre als docents personalitzar la seva ensenyança i als alumnes millorar el seu rendiment.
Les dades analitzades corresponen a la prova PIR de 2005. El procediment i els mètodes proposats són extrapolables a altres contexts educatius i de coneixements.

1 Introducció

Els tests de resposta múltiple s'utilitzen freqüentment en l'avaluació dels coneixements educatius. La puntuació que un alumne assoleix hauria de reflectir el nivell de coneixements que té sobre la matèria avaluada, però això no sempre és així. Altres aspectes poden fer esbiaixar la puntuació individual, invalidant les inferències que, a nivell individual, es poden extreure de les puntuacions. Un patró de resposta típic és aquell més freqüent en un determinat grup de referència o el que millor s'ajusta a un determinat model psicomètric. Els patrons contraris es consideren Patrons Atípics de Resposta (PAR) i la seva identificació permet detectar puntuacions que infraestimen o sobreestimen els coneixements.
En aquest estudi es proposa un mètode per identificar-ne la presència i tipologia en respostes a proves d'avaluació. Les dades analitzades corresponen a l'examen PIR de 2005. Es tracta d'una prova sumativa per a optar al títol d'especialista en psicologia clínica, i per tant, d'altes conseqüències per a les persones avaluades.

2 Models, supòsits i índexs

Hi ha una gran quantitat d'índexs per a detectar PAR. La majoria es basen en la Teoria de Resposta a l'ítem (fent servir models paramètrics o no paramètrics) o en la comparació amb patrons de respostes grupals. El supòsit que comparteixen tots aquests indexs és el d'unidimensionalitat, és a dir, que la prova només mesura un determinat coneixement i per tant la probabilitat de contestar correctament a una determinada pregunta depèn de les caracteréstiques psicomètriques d'aquesta i del nivell de la persona a la dimensió mesurada. Els models basats en la TRI assumeixen també la independència local, que implica que la resposta que una determinada persona doni a una pregunta no ha d'estar associada a las respostes que doni a la resta de les preguntes. En els models de TRI paramètrica el model ha de presentar un bon ajustament. Als models no paramètrics de TRI s'assumeix que les corbes característiques dels ítems són monòtones no decreixents, és a dir, que si una persona té un nivell més alt que una altra en el tret mesurat, no ha de tenir una probabilitat més baixa de donar una resposta correcta a l'ítem. Per a avaluar la presència de PAR s'han desenvolupat un elevat nombre d'índexs, però per a aquest estudi se n'han seleccionat tres, els que es consideren més eficients dins de cadascun dels tres contextos teòrics esmentats. Així s'han escollit els índexs HT, basat en referències grupals, $U3$, basat en la TRI no paramètrica i lz, basat en la TRI paramètrica.

3 Mètodes

Les dades analitzades corresponen a l'examen PIR de 2005 i van ser proporcionades pel Ministerio de Sanidad y Consumo. La base de dades incloïa 2057 files, una per a cada persona presentada a la prova, i 250 columnes, corresponents als ítems del test. De cada ítem es va tenir en compte si la resposta era incorrecta (codificada com a 0), correcta (codificada com a 1) o si no s'havia contestat (en blanc a la base de dades). Per a l'anàlisi es van considerar les no respostes com a respostes incorrectes.
La prova conté preguntes dels següents 9 continguts:

- Psicopatologia (54 ítems)
- Terapies i tractaments (41 ítems)
- Psicodiagnostic i avaluació conductual (21 ítems)
- Personalitat i psicologia diferencial (13 ítems)
- Processos bàsics i història (33 ítems)
- Psicometria, estadística i mètodes (20 ítems)
- Psicologia social i organitzacional (21 ítems)
- Psicologia evolutiva i educacional (25 ítems)
- Psicobiologia i psicofisiologia (22 ítems)

Les anàlisis dels patrons de respostes es van realitzar per a cadascuna d'aquestes subproves utilitzant el software R: El supòsit d'unidimensionalitat amb la funció `unidimTest` del paquet `ltm`, la independència local mitjançant una adaptació de la funció `MODFIT` del paquet `GGUM`, la monotonicitat no decreixent amb la funció `check.monotonicity` del paquet `mokken` i els índex de detecció de PAR amb el paquet `PerFit`. Els gràfics pel diagnòstic del PAR són d'elaboració pròpia, també programats amb R.
El procediment d'anàlisi va consistir en primer lloc en verificar els supòsits esmentats. En segon lloc es van calcular els índexs de detecció de PAR i, per a cadascun d'ells, es van considerar atípics els patrons de respostes amb valors més extrems, sent el punt de tall la mediana del punt corresponent al 5% més extrem de 1000 simulacions fetes a partir de les distribucions dels valors dels índexs.
Dels patrons atípics, identificats amb almenys un dels índexs, es van analitzar gràficament els de les persones que havien obtingut puntuacions globals més elevades. Per a fer-ho es va representar el percentatge de respostes correctes i incorrectes, esperades o inesperades, en tres parts de la prova: el terç d'ítems més fàcils, de dificultat mitjana i de més dificultat, comparant el perfil gràfic que proporcionen amb els que correspondrien a persones que han obtingut la mateixa puntuació però d'una forma tramposa, provant d'endevinar les preguntes que no saben, que han interpretat malament les preguntes més fàcils buscant-hi més dificultat de la que en realitat tenien (han estat çreatives") o que s'han despreocupat de les preguntes més fàcils, contestant-les incorrectament. Aquests perfils han estat simulats a partir de les dades i s'han comparat amb el patró observat mitjançant la distancia euclidiana.

4 Resultats

Respecte a la verificació dels supòsits, es van identificar 18 ítems que no complien l'assumpció de monotonicitat no decreixent. Aquests ítems podrien confondre els resultats i per aquest motiu es van eliminar. Totes les subproves van presentar bons indicadors d'unidimensionalitat bàsica i independència local.
De les 89 persones amb puntuacions globals més elevades, gairebé el 50% no presentava cap PAR en cap de les subproves. Pel que fa a la resta, el 40% presentava un en una de les àrees del test i el 10% restant en dues o tres.
A la Figura 1 s'hi presenta, com a exemple, el diagnòstic gràfic del cas 380 de la base de dades que presenta un PAR a l'àrea de Psicopatologia. La prova es divideix en tres parts equivalents en funció de la seva dificultat, representades pels tres rectangles de colors de 17, 17 i 16 ítems respectivament. En aquest cas, donat que la puntuació de la persona avaluada és de 46 la línia separa els 8 ítems més difícils, que haurien de tenir respostes incorrectes (rectangle exterior vermell), de la resta, que haurien de tenir respostes correctes (rectangle exterior verd). En cada àrea de la prova es mostren dos cercles proporcionals al seu nombre d'ítems, el color interior de ls quals indica el percentatge d'ítems contestats correctament (verd) o incorrectament (vermell).
A la Figura 2 s'hi presenten les gràfiques corresponents als patrons simulats sota diferents condicions de PAR mantenint la mateixa puntuació del cas analitzat. Les distàncies euclidianes entre els percentatges de respostes del perfil observat i de cada perfil simulat

indiquen que el perfil de respostes del cas 380 s'assembla més als perfils *creatiu* i *descuidat.*

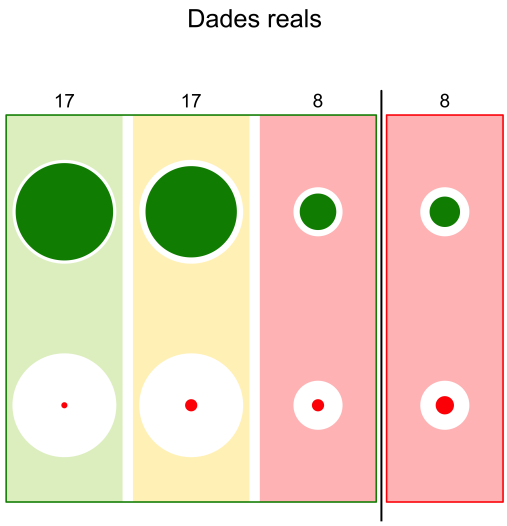


Figura 1: Perfil de respostes del cas 380 basat en el percentatge de respostes correctes i incorrectes (cercles verds i vermells) en cada part de la prova.

Aquest procediment analític s'ha repetit per a la resta de PAR amb puntuació més alta. Al voltant del 50% podrien ser classificats com a *creatius* o *descuidats*, el que implicaria, com en el cas presentat, que les seves puntuacions podrien subestimar el nivell de coneixement. Al voltant del 15% han estat classificats com a *endevinadors* i el 5% (1 cas) podria haver contestat de forma *tramposa*. Aquests perfils podrien sobrevalorar el coneixement. Finalment, gairebé el 30% podrien ser classificats com a normals, no coincidint amb el criteri dels índexs.

Figura 2: Perfils de respostes simulats, corresponents a la puntuació del cas 380. Els cercles mostren els percentatges de respostes correctes i incorrectes (cercles verds i vermells) en cada part de la prova.

5 Discussió

L'objectiu d'aquest projecte ha estat proposar un procediment d'anàlisi que permeti, no només detectar la presència de PAR si no també identificar-ne el tipus. El procediment proposat s'ha aplicat a una base de dades reals corresponents a 9 subproves en les que hem dividit l'examen PIR en funció del continguts dels seus ítems i ha combinat una metodologia analítica present a la literatura, amb una proposta gràfica pròpia.

La identificació de tipus de PAR ens ha permès dubtar de la validesa associada a les puntuacions de 4 persones que, per la seva puntuació, podrien haver obtingut una plaça com a psicòleg resident. Tot i que de forma relativa aquest número no sembla important, cal tenir en compte les conseqüències individuals que en aquesta prova té una avaluació no vàlida. Precisament per la importància d'aquestes conseqüències, els resultats d'aquestes

anàlisis s'han de prendre amb molta cautela, utilitzant-los com a screening i incorporant-hi després, si es pot, informació complementària.
Tot i que la prova avaluada tenia una finalitat sumativa, aquest procediment també pot ser útil en casos formatius. En aquests contexts, els docents podrien indagar en els motius subjacents als perfils que subestimen els coneixements avaluats, i utilitzar aquesta informació per a millorar el rendiment posterior d'aquests alumnes. Tot i que aquests procediments pot ser de gran utilitat en l'avaluació dels coneixements, cal seguir treballant-lo i millorant-lo.